Al-Ishlah: Jurnal Pendidikan
Vol.16, 4 (December, 2024), pp. 4363-4375
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v16i4.5671

Optimizing Question Quality in Junior High School Exams: Classical Test Theory Evaluation with ITEMAN 4.3

Agus Rohmatdi[1], Martono Martono[2], Zafrullah Zafrullah[3], Rina Safitri[4]

[1] Universitas Negeri Yogyakarta, Indonesia; agusrohmatdi.2020@student.uny.ac.id
[2] Universitas Negeri Yogyakarta, Indonesia; martono@uny.ac.id
[3] Universitas Negeri Yogyakarta, Indonesia; zafrullah.2022@student.uny.ac.id
[4] Universitas Negeri Yogyakarta, Indonesia; rina0045pasca.2021@student.uny.ac.id

ARTICLE INFO

Keywords:

Analysis of Question Items;
Classical Test Theory;
Junior High School

Article history:

Received 2024-07-13
Revised 2024-09-09
Accepted 2024-10-29

ABSTRACT

This research evaluates the psychometric properties of a test using Classical Test Theory, focusing on difficulty level, discrimination power, distractor effectiveness, and reliability with ITEMAN 4.3. Researchers analyzed 149 student answers taken using Cluster Random Sampling in the Arts and Culture subject at one of the junior high schools in Sleman, Yogyakarta, Indonesia. The analysis technique used is the level of difficulty, differential power, distractor effectiveness, and reliability. From the analysis of 20 questions, it can be concluded that this test shows an almost equal level of difficulty, with 12 questions at the medium level and eight questions at the easy level. The majority of questions have excellent discrimination, with their ability to effectively differentiate between students who understand the material well and those who do not. However, there is variation in the effectiveness of distractors, with some distractors being less effective in attracting the attention of students who do not understand the material. The reliability of this test is very good, with an Alpha value of 0.892, indicating high internal consistency in measuring the same concept. This research provides a comprehensive picture of the quality of evaluation instruments used in the context of measuring student abilities.

Corresponding Author:
Agus Rohmatdi
Universitas Negeri Yogyakarta, Indonesia; agusrohmatdi.2020@student.uny.ac.id

1. INTRODUCTION

Education forms a foundational pillar in human development, enabling individuals to cultivate character, refine skills, and expand their worldviews (Beghetto, 2023; Ramadhani & Retnawati, 2024). Beyond academic achievement, education fosters moral and ethical growth, nurturing individuals with integrity and a strong sense of societal responsibility (Holden et al., 2021; Poitras Pratt & Gladue, 2022). This process is deeply interwoven with broader social, economic, and cultural factors, creating a dynamic ecosystem that shapes individuals and communities alike (Mansilla & Wilson, 2020; Purwani & Arvianti, 2020; Sulastri, 2023). High-quality education is essential for fostering an intelligent, critical, and competitive society, equipping individuals to contribute meaningfully to national development and global progress (Rumawatine, 2023; Zafrullah et al., 2023). Schools, as core institutions within the

educational system, play a crucial role in this endeavor, and the quality of exams administered is pivotal to assessing and enhancing student learning outcomes. In this context, optimizing question quality in junior high school exams is fundamental to ensuring accurate assessments. By applying Classical Test Theory (CTT) and leveraging tools like ITEMAN 4.3, educators can rigorously evaluate and improve test item quality, creating fair and reliable assessments that support meaningful educational advancement.

Schools are a key element in the education sector, which plays an important role in shaping the nation's next generation (Hasanah et al., 2023; Zafrullah et al., 2024). Apart from providing academic knowledge, schools also shape students' character, social skills and moral values. School quality is the main factor in ensuring that every student gets the best education that suits their needs and potential (Chankseliani et al., 2021; Tien et al., 2022; Ulwiyah, 2023). Through these institutions, children gain guidance, knowledge, and experience that help them develop into competent and responsible individuals (Firmansyah et al., 2024; Susmayati et al., 2024; Zafrullah & Zetriuslita, 2021). Therefore, attention to the quality and accessibility of schools must continue to be increased so that all levels of society can enjoy the benefits of quality education (Høibo et al., 2024; Okada et al., 2024; Zafrullah et al., 2024). Apart from that, it is also important to create a conducive learning environment and provide adequate educational resources, so that every student has the same opportunity to achieve success. To achieve these results, it is important for schools to continue to flow the learning process and student development regularly. This evaluation functions as a tool to measure the effectiveness of the education provided and as a reference in improving teaching methods, so that each student can reach their maximum potential.

Evaluation at the end of learning is a crucial stage in the educational process, because through this evaluation, students' progress and understanding of the material being taught can be measured objectively (Fan & Zhong, 2022; Sewang & Halik, 2020). Evaluation at the end of learning has an important function in assessing the effectiveness of the teaching methods used and evaluating the extent to which learning objectives have been achieved (Chang et al., 2024; Rasul et al., 2023). Positive end-of-learning evaluations can provide constructive feedback for students, helping them identify strengths and areas for improvement (Rumahlewang et al., 2023; Schellekens et al., 2021). Apart from that, evaluation at the end of learning is also important for teachers and educational institutions in designing better teaching strategies in the future. Thus, this evaluation not only functions as a measurement tool, but also as a means to continue to improve the quality of education and ensure that each student receives optimal education according to his abilities and potential (Sokhanvar et al., 2021). One type of evaluation that is often used is using multiple choice questions.

Multiple choice is a form of evaluation that is often used at various levels of education because of its ease in preparation and assessment (Akimov & Malin, 2020; Guangul et al., 2020). Multiple choice is an effective tool for measuring students' understanding of the material that has been taught, because it is able to cover various aspects of knowledge in one set of questions (Harahap, 2024; Immanuel et al., 2024; Pramana et al., 2024). In addition, multiple choice can test students' ability to think critically and make the right decisions in a relatively short time. Multiple choice is important in providing a general idea of the extent to which students have mastered certain concepts, as well as assisting teachers in identifying topics that may require further explanation (Tian et al., 2024; Zhang et al., 2024). Thus, although multiple choice is not the only evaluation method, its existence still plays a vital role in the education system, especially in providing fast and accurate feedback regarding student academic achievement (Kubiszyn & Borich, 2024). There are two types of multiple-choice question assessment, namely Classical Test Theory (CTT) and item response theory (IRT). Classical Test Theory focuses on analyzing the number of correct and incorrect answers and the overall difficulty of the questions, while Item Response Theory goes more in-depth by considering the characteristics of the individuals answering the questions, such as their level of ability in solving certain questions. In this study, researchers focused on classical test theory.

Classical test theory is an evaluation method that focuses on simple statistical analysis of exam results, including calculating the number of correct and incorrect answers and the overall level of difficulty of the questions (Baharuddin et al., 2024; Yu, 2020). This theory has long been used in the world of education to evaluate student learning outcomes and measure the level of teaching effectiveness.

Classical test theory plays an important role in providing a general idea of how well students understand the material being taught and the extent to which they can apply it in exam situations (Ananda & Pratama, 2024). Classical test theory remains relevant as a basic evaluation tool that helps assess and improve the learning process. However, few people use classical test theory today because not many people understand its complexity and limitations in providing accurate in-depth information about individual abilities.

The benefit of classical test theory is the ability to provide a simple but effective analysis of student exam results, including calculating correct and incorrect answers and the level of difficulty of the questions. The use of classical test theory allows educators to assess students' understanding of the material and the effectiveness of teaching in an easily accessible way. Thus, classical test theory plays an important role in educational evaluation because it provides useful information for improving the learning process and adapting teaching strategies according to student needs. One application that can help analyze questions on classical test theory is ITEMAN.

The ITEMAN application is software used to carry out test analysis based on classical test theory, this is proven by several studies involving ITEMAN as a data analysis application (Ningsih & Istiyono, 2023; Triono et al., 2020). ITEMAN is important because it makes it easy to calculate basic statistics such as reliability, difficulty of questions, and distinguishing power of questions in a test (Hanifah, 2023; Wahab et al., 2023). ITEMAN advantages include ease of data input using Microsoft Excel, analysis results that are easy to understand, as well as an attractive appearance and various versions (Hodiyanto & Saputro, 2018). With these features, ITEMAN is very beneficial for teachers because it simplifies the question evaluation process, allows them to analyze question performance effectively and understand the results clearly. This helps teachers organize and adapt questions to better suit students' abilities, thereby improving the quality of evaluation and teaching effectiveness. By using ITEMAN, users can identify the weaknesses and strengths of each question in the test, as well as analyze the level of difficulty and effectiveness in measuring students' understanding of the subject matter. Therefore, ITEMAN becomes a practical application for teachers and researchers in developing better tests and improving the overall quality of educational evaluation.

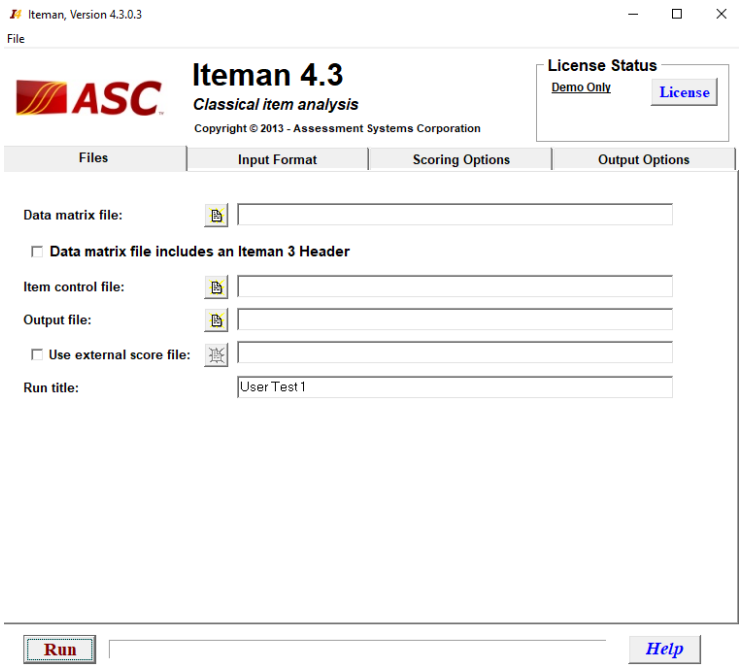

Figure 1. Initial Display of ITEMAN 4.3

Findings from observations at one of the junior high schools in Sleman, Yogyakarta, Indonesia, show that art and culture subject teachers cannot yet know the exact qualities of their students. Therefore, it is important to evaluate the quality of the questions in order to obtain a clearer picture of students' abilities

and understanding of arts and culture material. The results of previous research also show that ITEMAN analysis can help teachers evaluate question performance and student performance (Himawan & Nurgiyantoro, 2022; Kustati & Amelia, 2023). By carrying out this analysis, it is hoped that this research can provide important implications, namely as an evaluation for teachers in creating more effective questions. This will help teachers prepare questions that are more appropriate to students' abilities so that they can increase students' understanding and achievement of learning outcomes in arts and culture subjects.

2. METHODS

This research is descriptive quantitative research that focuses on questions using the ITEMAN 4.3 application. Quantitative research is a research method that uses an approach to collect and analyze numerical or quantitative data, which in this context is used to measure various parameters (Mertens, 2023; Mohajan, 2020; Strijker et al., 2020). In this research, the parameters measured include the level of difficulty, differentiation, and reliability of the 8th grade Arts and Culture exam questions in junior high schools in Sleman, Yogyakarta, Indonesia, in 2024. Research parameters such as difficulty, discrimination, reliability, and distractor effectiveness are important for evaluating the quality of test questions because each parameter provides in-depth information about the extent to which the questions can accurately measure students' abilities, distinguish between different levels of understanding, and ensure that the questions have consistency and is not misleading. Thus, this research aims to provide a clear and measurable picture of the quality of the questions used in the test.

Data collection was carried out using tests that had been prepared previously and given to 149 students who were the research samples determined by Cluster Random Sampling. Students answer the test according to the instructions given, and the results are used as data for further analysis. The data analysis techniques used include evaluating the level of difficulty of the questions to find out how difficult the questions are, measuring the power of differentiation to assess how well the questions can differentiate between students with different abilities, evaluating the effectiveness of distractors to identify the most interesting answer choices other than the correct answer, as well as calculations. Reliability is to determine how consistent the test results are if repeated measurements are made. The criteria for levels of difficulty and different strengths can be seen in Table 1 and Table 2. To evaluate the criteria for distractor effectiveness, ITEMAN 4.3 uses the proportional endorsement method, which helps in assessing how often answer choices other than the correct one are chosen by students. Meanwhile, to ensure good reliability, the expected reliability value is above 0.70, in accordance with the standards suggested by Cho & Kim (2015).

Good questions are questions with a difficulty level that is in the range of 30% to 70% so that they are not too easy or difficult for students (Bano, 2023). The question must also have a differential power of at least 0.30, which shows its ability to differentiate between students who have good understanding and those who have less. In addition, the effectiveness of the distractor must be high, that is, the incorrect answer option must be convincing enough to test the student's understanding and reduce the possibility of random guessing (Amalia & Widayati, 2012).

Table 1. Difficulty Level Classification

Difficulty Level	Description
0.00 – 0.30	Hard
0.31 – 0.70	Medium
0.71 – 1.00	Easy

Source: Istiyono (2020)

Table 2. Discrimination Power Criteria

Criteria	Description
$D \leq 0.1999$	Bad
$0.200 - 0.299$	Fairly good (Needs Revision)
$0.300 - 0.399$	Medium (No need to revise)
$D \geq 0.400$	Very Good

Source: Istiyono (2020)

3. FINDINGS AND DISCUSSION

The aim of this analysis is to assess the quality of questions in the 8th grade Arts and Culture subject using ITEMAN 4.3 software, which allows detailed evaluation of various parameters such as difficulty level, different power, distractor effectiveness, and question reliability. First, the author interprets the Summary Statistics in Table 3.

Table 3. Summary Statistics on the Results of Question Item Analysis with ITEMAN 4.3

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	20	13.919	5.133	2	20	0.696	0.505

Source: ITEMAN 4.3, Data analyzed on June 23, 2024

The results of item analysis show that of the 20 questions tested, the average score obtained by students was 13.919 with a Standard Deviation (SD) of 5.133. This high Standard Deviation indicates significant variation in student scores around the mean, meaning there are large differences in students' understanding and ability of the material tested. A large Standard Deviation indicates that there are students who get very low scores, such as 2, and there are also students who get the highest score, namely 20. This reflects a gap in student mastery of the material, where some students do not master the material at all. , while others are able to master the material very well.

In addition, the average value of the proportion of correct answers (Mean P) is 0.696, which means that around 69.6% of all questions were answered correctly by the average student. This shows that most of the questions are at a medium level of difficulty. The average point-biserial correlation (Mean r_{pbis}) is 0.505, indicating that the items have good discrimination and are able to differentiate between students who have a high and low understanding of the material. These results indicate that these questions are quite effective in measuring students' abilities accurately and can be relied upon as learning evaluation instruments in Arts and Culture subjects.

3.1 Difficulty Levels

Researchers conduct difficulty level analysis because it is important to know how difficult or easy the questions are for students. Difficulty levels play a role in ensuring that the questions are appropriate to student abilities, assist in identifying areas that need improvement in teaching, and ensure that each question effectively measures student understanding.

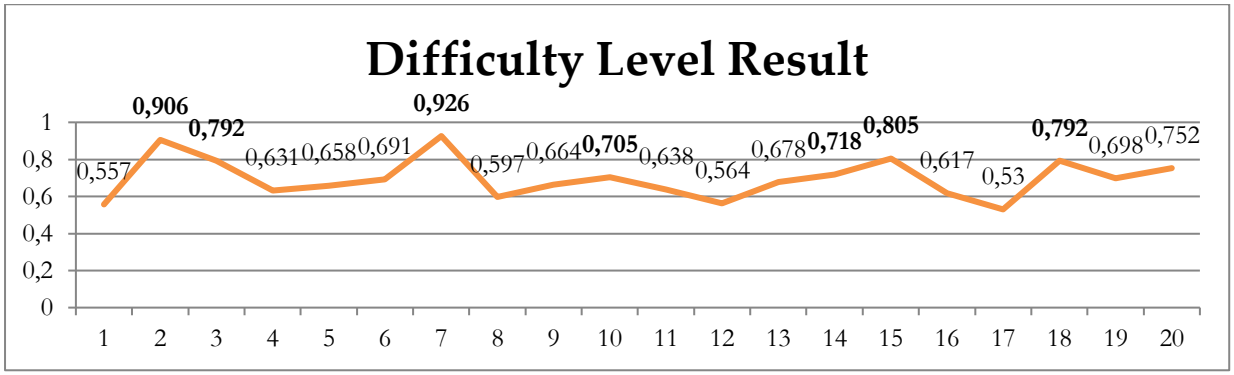

Figure 1. Chart Regarding Difficulty Levels for 20 Numbers analyzed with ITEMAN 4.3

The results of the difficulty level analysis show that of the 20 questions analyzed, the majority of the questions are in the medium and easy categories. There are 12 questions that are categorized as medium, namely questions number 1, 4, 5, 6, 8, 9, 11, 12, 13, 16, 17, and 19, with a difficulty level ranging from 0.530 to 0.698. This shows that most of the questions have a medium level of difficulty and can be answered well by most students.

Apart from that, there are 8 questions that fall into the easy category, namely questions number 2, 3, 7, 10, 14, 15, 18, and 20, with a difficulty level between 0.705 to 0.926. These questions showed that they were easier for students, with a large proportion of students being able to answer them correctly. The overall distribution of difficulty levels shows variation in question difficulty, with more questions being at medium difficulty compared to easy questions.

3.2 Different Power

The difference power analyzed by researchers aims to measure the extent to which each question item is able to differentiate between students who have high and low understanding of the subject matter. This analysis is important to ensure that the questions used in the test can effectively highlight differences in student ability levels, so that the test results can provide an accurate picture of the distribution of abilities in the group of students being tested.

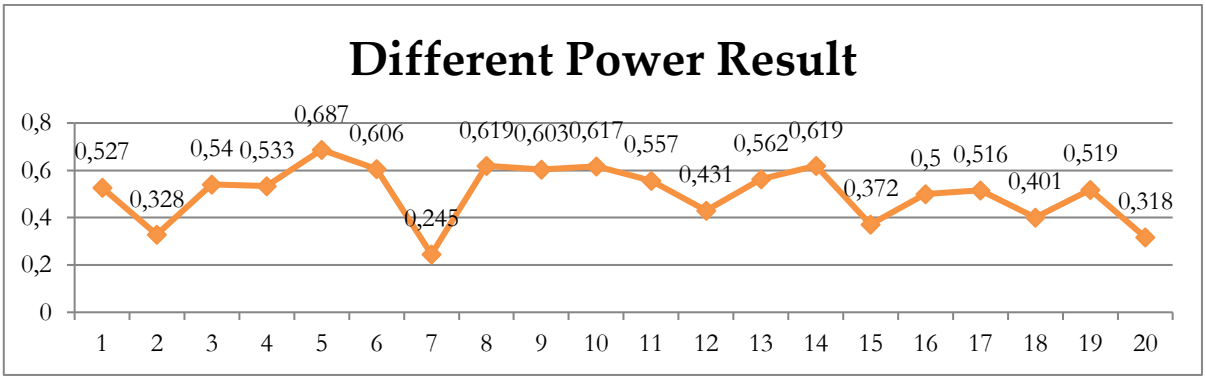

Figure 2. Chart of Differential Power Results on 20 Questions Analyzed with ITEMAN 4.3

The results of the differential power analysis showed that of the 20 questions analyzed, the majority of questions had very good differential power (Very Good). The questions with very good differential power range from 0.401 to 0.687, which includes most of the questions, such as questions number 1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, and 19. This indicates that these question items are very effective in distinguishing students with good and poor understanding of the subject matter.

Apart from that, there are several questions that have different strengths in the medium and fairly good categories. The questions with moderate power differences are questions number 2, 15, and 20, with values ranging from 0.318 to 0.372. Meanwhile, the question with a fairly good difference is question number 7, with a value of 0.245. Although these questions still function in differentiating students' levels of understanding, their effectiveness is not as strong as questions that have very good discrimination. This analysis is important to ensure that all questions in the test can effectively assess and differentiate student abilities.

3.3 Distractor Effectiveness

Due to the limited number of words, the researcher only took one question each at a medium level of difficulty, and it was easy to analyze the effectiveness of the distractors. This aims to obtain a representative picture of how each answer choice (distractor) functions in two different difficulty categories, so as to provide more focused and in-depth insight into the quality and effectiveness of questions in measuring student understanding.

Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
7	ITEM07	A	Yes	4	1	

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
149	0.926	0.245	0.457	0.893

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	138	0.926	**0.245**	0.457	14.341	5.041	Maroon	**KEY**
B	3	0.020	**-0.133**	-0.383	8.333	5.859	Green	
C	1	0.007	**-0.147**	-0.639	4.000	0.000	Blue	
D	7	0.047	**-0.157**	-0.339	9.429	4.650	Olive	
Omit	0							
Not Admin	0							

Figure 3. Details of Wrong Questions in the "Easy" Category in ITEMAN 4.3

Analysis of the distractor's effectiveness on questions in the easiest category shows that option A is the key to the correct answer, with an answer proportion of 0.926 and a point-biserial correlation (r_{pbis}) of 0.245. A positive r_{pbis} value for option A indicates that students who choose this answer tend to have a higher overall score, indicating that this answer is effective in differentiating between students who understand the material well and those who do not. In contrast, the other distractor options (B, C, and D) had negative r_{pbis} values (-0.133, -0.147, and -0.157), meaning students who chose these answers tended to have lower overall scores. The proportion of answers choosing options B, C, and D was very low (0.020, 0.007, and 0.047), indicating that these distractors were less effective because they were rarely chosen by students. This indicates that the distractor option was not successful in attracting students who did not understand the material.

Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
17	ITEM17	C	Yes	4	1	

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
149	0.530	0.516	0.648	0.887

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	14	0.094	**-0.379**	-0.659	7.714	2.164	Maroon	
B	21	0.141	**-0.408**	-0.636	8.524	3.995	Green	
C	79	0.530	**0.516**	0.648	16.734	3.773	Blue	**KEY**
D	35	0.235	**-0.012**	-0.016	13.286	4.528	Olive	
Omit	0							
Not Admin	0							

Figure 4. Details of Wrong Questions in the "Medium" Category in ITEMAN 4.3

Analysis of the distractor's effectiveness in one of the medium category questions shows that option C is the key to the correct answer, with an answer proportion of 0.530 and a point-biserial correlation (r_{pbis}) of 0.516. A high r_{pbis} value in option C indicates that students who chose this answer had a higher overall score, indicating that this answer effectively differentiates between students who understand the material well and those who do not. In contrast, the other distractor options (A, B, and

D) had negative r_{pbis} values (-0.379, -0.408, and -0.012), meaning students who chose these answers tended to have lower overall scores. The proportion of responses selecting options A, B, and D (0.094, 0.141, and 0.235) indicates that although some students were interested in these distractors, they were less effective in separating students based on their level of understanding.

When compared with the results in the previous easy question category, it can be seen that both questions show a similar pattern where the answer key has a significant positive r_{pbis}, indicating effectiveness in differentiating student abilities. However, in the previous easy question, the proportion of students who chose the correct answer (0.926) was much higher than in this question (0.530), indicating that this question was more difficult. Additionally, the distractors on these more difficult questions (A, B, and D) had larger negative r_{pbis} values and more evenly distributed proportions than the previous easy questions, indicating greater variation in student choices and the need for revision to increase effectiveness these distractors in assessing student understanding accurately.

Based on the results of the distractor effectiveness analysis, there are several answer options with negative results. So, it can be suggested to improve the quality of distractors in the easiest category questions by increasing the attractiveness and relevance of incorrect answer options. A distractor with a negative r_{pbis} value and a very low proportion of answers indicates that the option is not effective in testing student understanding. To increase the effectiveness of distraction, the wrong option should be designed to be more convincing and relevant to the material being tested, so that it can attract the attention of students who do not understand the material and differentiate them from students who really master the material. Thus, effective distractors can help improve the ability of questions to measure students' level of understanding more accurately.

3.4 Reliability Results

Apart from the effectiveness of distractors, the author also analyzes reliability, which aims to ensure the consistency and reliability of test results when measuring students' abilities repeatedly.

Table 3. Overall Reliability Results on Question Items analyzed with ITEMAN 4.3

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	0.892	1.684	0.816	0.729	0.809	0.899	0.844	0.894

Source: ITEMAN 4.3, Data analyzed on June 23, 2024

The findings of this study reinforce and expand upon prior research on the psychometric properties of multiple-choice assessments in educational settings. Notably, studies such as those conducted by Kastrara et al. (2024) have demonstrated that many multiple-choice tests, particularly in subjects like Citizenship Education (PKn), exhibit a medium level of difficulty, with overall discrimination power in the medium range and good reliability, as indicated by an Alpha value of 0.854. Such psychometric properties align with the expected performance of assessments based on Classical Test Theory (CTT), which emphasizes the need for balanced levels of difficulty and high discrimination power to distinguish student ability effectively (Fan, 1998; Crocker & Algina, 2006). The present study, focused on the Arts and Culture subject for junior high school students, similarly shows a balanced level of difficulty across items, with 12 questions categorized as medium and eight as easy, supporting the assertion that well-structured assessments achieve a range that accommodates varying student abilities. Furthermore, the discrimination power observed in this study aligns with high expectations for effective assessments, where most items effectively differentiate between students who understand the material well and those who do not. This alignment with previous findings underscores the reliability and appropriateness of CTT methods in assessing junior high school exams (DeMars, 2010).

A crucial aspect of the study is the examination of distractor effectiveness, a feature often underexplored in test evaluations. While the overall discrimination was strong, indicating effective differentiation among students, there was notable variability in distractor effectiveness. Research indicates that distractors play a key role in enhancing item validity and reliability by challenging students' comprehension and discouraging guesswork (Tarrant, Ware, & Mohammed, 2009; Downing, 2006). However, in this study, some distractors failed to capture the attention of students who did not fully understand the material, which could limit the assessment's capacity to measure nuanced levels of student comprehension. These findings resonate with prior research that has highlighted distractor quality as an area for improvement in many educational assessments (Haladyna & Downing, 1993). Ineffective distractors can skew student performance data, suggesting a need for careful review and revision of items where distractors do not fulfill their intended purpose. Improving distractor quality could help ensure that items more accurately reflect a student's understanding and avoid common pitfalls, such as overly predictable or unengaging answer options (Rodriguez, 2005).

The study's reliability measure, with an Alpha value of 0.892, demonstrates high internal consistency, suggesting that the test items are well-aligned in assessing the same underlying concept of student understanding in the Arts and Culture subject. This result indicates that the test is a cohesive tool for measuring student knowledge and skills within the targeted content area (Taber, 2018). In comparison to other studies, where reliability scores hover in the lower 0.8 range, this higher reliability indicates that the test may provide more stable and consistent results over repeated administrations. Such high reliability is essential in educational testing, as it helps educators make informed decisions based on assessment outcomes that are genuinely reflective of student capabilities (Field, 2013).

This research underscores the significance of applying CTT methods, supported by software like ITEMAN 4.3, in refining educational assessments to meet rigorous psychometric standards (Ruch, 2014). For future studies, exploring the potential of item response theory (IRT) as a complementary approach to CTT could provide additional insights, particularly in understanding the functionality of distractors and enhancing the discrimination power of test items (Embretson & Reise, 2000). Moreover, expanding the research scope to include other subjects and diverse school environments could offer a broader understanding of test item quality across various educational contexts. By prioritizing the continuous improvement of question quality, especially in terms of distractor effectiveness, educators can design assessments that more accurately measure and support student learning and development (Downing & Haladyna, 2006).

4. CONCLUSION

This research aims to analyze the quality of exam questions for class VIII Arts and Culture subjects in junior high schools in Sleman, Yogyakarta, using the ITEMAN 4.3 application to measure the difficulty, differentiation, reliability, and effectiveness of distractors. Research findings show that the test has a difficulty level that is almost balanced between the medium and easy categories, good discrimination, and very high reliability with an Alpha value of 0.892. However, there is variation in the effectiveness of distractors, with some distractors being less effective in attracting students' attention. The educational implications of these findings suggest that although these tests are generally reliable and accurate, improvements in distractor effectiveness are needed to improve test quality and ensure fairer evaluations. Recommendations for improvement include item revisions to improve distractor quality and further testing to ensure that each item can better differentiate between students with different understandings. Limitations of this study include the sample being limited to one geographic location and certain types of subjects, so the results may not fully represent conditions in other areas or subjects. For future research, it is recommended to expand the scope of research to include different subjects and locations, as well as use additional evaluation methods to obtain a more comprehensive picture of the quality of exam questions.

REFERENCES

Akimov, A., & Malin, M. (2020). When old becomes new: a case study of oral examination as an online assessment tool. *Assessment & Evaluation in Higher Education*, *45*(8), 1205–1221.

Amalia, A. N., & Widayati, A. (2012). Analisis butir soal tes kendali mutu kelas XII SMA mata pelajaran ekonomi akuntansi di kota Yogyakarta tahun 2012. *Jurnal Pendidikan Akuntansi Indonesia*, *10*(1).

Ananda, R., & Pratama, F. F. (2024). Classic Learning Media Such As Image Media: Do They Still Have An Impact On Learning In Elementary Schools? *International Journal of Education and Teaching Zone*, *3*(2), 196–209.

Baharuddin, B., Handayani, L., & Rusli, R. (2024). *Enhancing Biochemistry Assessment Quality in Medical Education Through Item Response Theory (IRT)*.

Bano, V. O. (2023). Perbandingan Hasil Belajar Peserta Didik Berdasarkan Tingkat Kesukaran, Daya Beda Dan Efektivitas Pengecoh Pada Butir Soal Di Sman 1 Pandawai. *Jurnal Edusavana*, *1*(1), 30–41.

Beghetto, R. A. (2023). Broadening horizons of the possible in education. *Possibility Studies & Society*, *1*(4), 414–426.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., & Wang, Y. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45.

Chankseliani, M., Qoraboyev, I., & Gimranova, D. (2021). Higher education contributing to local, national, and global development: new empirical and conceptual insights. *Higher Education*, *81*(1), 109–127.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230.

Crocker, L., & Algina, J. (2006). Introduction to classical and modern test theory. Wadsworth.

DeMars, C. E. (2010). Item response theory. Oxford University Press.

Downing, S. M. (2006). Twelve steps for effective test development. Medical Teacher, 28(7), 608-614. https://doi.org/10.1080/01421590600638092

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Routledge.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58(3), 357-381. https://doi.org/10.1177/0013164498058003001

Fan, X., & Zhong, X. (2022). Artificial intelligence-based creative thinking skill analysis model using human–computer interaction in art design teaching. *Computers and Electrical Engineering*, *100*, 107957.

Firmansyah, M. D., Sugihartini, D. P., & Rachman, I. F. (2024). Transformasi Pendidikan Melalui Kolaborasi Pemerintah, Swasta, Dan Masyarakat Untuk Literasi Digital Demi Pembangunan Berkelanjutan 2030. *MERDEKA: Jurnal Ilmiah Multidisiplin*, *1*(4), 317–327.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

Guangul, F. M., Suhail, A. H., Khalit, M. I., & Khidhir, B. A. (2020). Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educational Assessment, Evaluation and Accountability*, *32*, 519–535.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? Educational and Psychological Measurement, 53(4), 999-1010. https://doi.org/10.1177/0013164493053004013

Hanifah, F. (2023). Analysis of Final Semester Assessment Question Items for Class XI Indonesian History Subjects. *Indonesian Journal of History Education*, *8*(2), 89–106.

Harahap, A. (2024). *Evaluasi Pembelajaran Berbasis Hots Dalam Kurikulum Merdeka*. Penerbit Adab.

Hasanah, U., Hakim, I. U., & Zain, Z. F. S. (2023). Islamic Education in the Society 5.0 Era: Lesson to

Learn. *IJECA (International Journal of Education and Curriculum Application)*, *6*(1), 21–32.
Himawan, R., & Nurgiyantoro, B. (2022). Analisis butir soal latihan penilaian akhir semester ganjil mata pelajaran bahasa Indonesia kelas VIII SMPN 1 Bambanglipuro Bantul menggunakan program ITEMAN. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, Dan Pengajarannya*, *8*(1), 160–180.
Hodiyanto, H., & Saputro, M. (2018). Workshop pembuatan dan analisis butir soal menggunakan Iteman pada Madrasah Aliyah Miftahul Huda Kecamatan Sungai Ambawang. *Transformasi: Jurnal Pengabdian Masyarakat*, *14*(2), 85–90.
Høibo, I. H., Seitamaa-Hakkarainen, P., & Groth, C. (2024). Teachers' pedagogical beliefs in Norwegian school makerspaces. *International Journal of Technology and Design Education*, 1–18.
Holden, O. L., Norris, M. E., & Kuhlmeier, V. A. (2021). Academic integrity in online assessment: A research review. *Frontiers in Education*, *6*, 639814.
Immanuel, C., Manik, S. D. P., Nababan, A. P., Sianturi, B. Y., & Hasnah, A. (2024). Analisis Butir Soal Pilihan Ganda Mata Pelajaran Matematika Materi Bangun Datar Berbasis Budaya Lokal di SDN 106163 Bandar Klippa. *Jurnal Ilmiah Multidisiplin Terpadu*, *8*(6).
Istiyono, E. (2020). Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern. *Yogyakarta: UNY Press. L, I.(2019). Evaluasi Dalam Proses Pembelajaran. Jurnal Manajemen Pendidikan Islam*, *9*, 478–492.
Kastrara, R., Riantoro, E. S., & Bakti, A. A. (2024). Analisis Butir Soal Dengan Iteman 4.0 Pada Penilaian Akhir Semester Sekolah Dasar. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *9*(1), 5813–5823.
Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.
Kustati, M., & Amelia, R. (2023). Analisis Kualitas Butir-Butir Soal Ulangan Harian Pendidikan Agama Islam Dengan Menggunakan Iteman. *SOKO GURU: Jurnal Ilmu Pendidikan*, *3*(3), 142–155.
Mansilla, V. B., & Wilson, D. (2020). What is global competence, and what might it look like in Chinese schools? *Journal of Research in International Education*, *19*(1), 3–22.
Mertens, D. M. (2023). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage publications.
Mohajan, H. K. (2020). Quantitative research: A successful investigation in natural and social sciences. *Journal of Economic Development, Environment and People*, *9*(4), 50–79.
Ningsih, S. N., & Istiyono, E. (2023). An Application of Classical Test Theory for Item Characteristic Analysis of Chemical Literacy Instruments. *Jurnal Pendidikan Kimia Indonesia*, *7*(2), 58–68.
Okada, A., Panselinas, G., Bizoi, M., Malagrida, R., & Torres, P. L. (2024). Fostering transversal skills through open schooling with the CARE-KNOW-DO framework for sustainable education. *Sustainability*, *16*(7), 2794.
Poitras Pratt, Y., & Gladue, K. (2022). Re-defining academic integrity: Embracing Indigenous truths. In *Academic integrity in Canada: An enduring and essential challenge* (pp. 103–123). Springer International Publishing Cham.
Pramana, R., Melissa, C., Rivaldo, I., Ariska, W., Mahisa, A., & Eka, R. (2024). Analisis Data Profil SMA Dalam Kolaborasi Keterampilan Mengajar Mahasiswa Geografi. *SOSIAL: Jurnal Ilmiah Pendidikan IPS*, *2*(2), 187–200.
Purwani, T., & Arvianti, I. (2020). The Economic Empowerment Model of Multicultural Society. *The 2nd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2020)*, 171–178.
Ramadhani, A. M., & Retnawati, H. (2024). Computational Thinking and its Application in School: A Bibliometric Analysis (2008-2023). *International Conference on Current Issues in Education (ICCIE 2023)*, 329–338.
Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), 41–56.
Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice, 24(2), 3-13.

https://doi.org/10.1111/j.1745-3992.2005.00006.x

Ruch, F. L. (2014). Principles of Educational and Psychological Measurement. McGraw-Hill.

Rumahlewang, E., Pelenkahu, N., Ali, M. I., & Tatipang, D. P. (2023). *Formative Assesment*. Cattleya Darmaya Fortuna.

Rumawatine, Z. (2023). The Effect Of Personal Learning Models On Self-Confidence And Learning Outcomes To Play Soccer In Extracurricular Men's Soccer. *JIM: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, *8*(2), 864–873.

Schellekens, L. H., Bok, H. G. J., de Jong, L. H., van der Schaaf, M. F., Kremer, W. D. J., & van der Vleuten, C. P. M. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation*, *71*, 101094.

Sewang, A., & Halik, A. (2020). Learning Management Model of Islamic Education based on Problem: A Case Study of the Tarbiyah and Adab Department of IAIN Parepare. *Talent Development & Excellence*, *12*(1), 2731–2747.

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, *70*, 101030.

Strijker, D., Bosworth, G., & Bouter, G. (2020). Research methods in rural studies: Qualitative, quantitative and mixed methods. *Journal of Rural Studies*, *78*, 262–270.

Sulastri, S. (2023). Application of the Assignment Method in Enhancing Student Learning Enthusiasm in the Subject of Jurisprudence. *Elementaria: Journal of Educational Research*, *1*(1 SE-Articles), 54–64. https://doi.org/10.61166/elm.v1i1.5

Susmayati, S., Veranty, A., Cahyani, L. U., Rambe, S. M., Jahra, S. S., & Safitri, R. (2024). Mempertahankan Jati Diri Identitas Nasional Di Era Globalisasi Dan Digitalilasi. *JURNAL TIPS JURNAL RISET, PENDIDIKAN DAN ILMU SOSIAL*, *1*(1), 62–70.

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Medical Education, 9(1), 1-8. https://doi.org/10.1186/1472-6920-9-40

Tian, P., Fan, Y., Sun, D., & Li, Y. (2024). Evaluating students' computation skills in learning amount of substance based on SOLO taxonomy in secondary schools. *International Journal of Science Education*, 1–23.

Tien, N. H., Ngoc, N. M., Trang, T. T. T., & Mai, N. P. (2022). Sustainable Development of Higher Education Institutions in Developing Countries: Comparative Analysis of Poland and Vietnam. *Contemporary Economics*, *16*(2).

Triono, D., Sarno, R., & Sungkono, K. R. (2020). Item Analysis for Examination Test in the Postgraduate Student's Selection with Classical Test Theory and Rasch Measurement Model. *2020 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 523–529.

Ulwiyah, S. (2023). Rasch Model Analysis On Mathematics Test Instruments: Biblioshiny (1983-2023). *Mathematics Research and Education Journal*, *7*(2), 1–13.

Wahab, A., Hasibuan, A., Siregar, R., & Ningsih, T. Z. (2023). Item Analysis of Final Examination Questions for Social Studies in Junior High Schools through the ITEMAN Program. *Journal of Education Research and Evaluation*, *7*(3).

Yu, C. H. (2020). Objective measurement: How Rasch modeling can simplify and enhance your assessment. *Rasch Measurement: Applications in Quantitative Educational Research*, 47–73.

Zafrullah, Z., Hakim, M. L., & Angga, M. (2023). ChatGPT open AI: Analysis of mathematics education students learning interest. *Journal of Technology Global*, *1*(01), 1–10.

Zafrullah, Z., Sultan, J., Ayuni, R. T., & Uleng, A. T. (2024). Analisis Kemandirian Belajar Matematika

Siswa Berdasarkan Gender dan Aspek di Sekolah Menengah Atas. *Perspektif Pendidikan Dan Keguruan*, *15*(1), 29–38.

Zafrullah, Z., & Zetriuslita, Z. (2021). Learning interest of seventh grade students towards mathematics learning media assisted by Adobe Flash CS6. *Math Didactic: Jurnal Pendidikan Matematika*, *7*(2), 114–123.

Zhang, H., Perry, A., & Lee, I. (2024). Developing and Validating the Artificial Intelligence Literacy Concept Inventory: an Instrument to Assess Artificial Intelligence Literacy among Middle School Students. *International Journal of Artificial Intelligence in Education*, 1–41.