Journal of Educational Research and Evaluation
Volume 7, Issue 3, 2023, pp. 526-536
P-ISSN: 2597-422x E-ISSN: 2549-2675
Open Access: https://doi.org/10.23887/jere.v7i3.59239

Item Analysis of Final Examination Questions for Social Studies in Junior High Schools through the ITEMAN Program

Abdul Wahab[1*], Annim Hasibuan[2], Roswani Siregar[3], Risnawaty[4], Tri Zahra Ningsih[5]

[1]Universitas Muslim Indonesia, Makassar, Indonesia
[2]Universitas Islam Labuhan Batu, Labuhanbatu, Indonesia
[3]Universitas Al-Azhar Medan, Medan, Indonesia
[4]UMN Al-Washiyah, Medan, Indonesia
[5]SMP Negeri 46 Kerinci, Kerinci, Indonesia

ARTICLE INFO

Article history:
Received March 02, 2023
Revised March 08, 2023
Accepted July 10, 2023
Available online August 25, 2023

Kata Kunci :
Analisis butir soal, IPS, ITEMAN

Keywords:
Item Analysis, Social Learning, ITEMAN.

ABSTRAK

Analisis kulaitas butir tes sangat penting dilakukan, karena akan mempengaruhi hasil tes itu sendiri. Permasalahannya seringkali guru tidak melakukan analisis butir tes, sehingga hasil tes kurang tepat menggambarkan potensi peserta didik. Untuk itu, penelitian ini bertujuan untuk menganalisis butir soal ujian akhir semester mata pelajaran IPS kelas VII. Jenis penelitian ini adalah deskripsi kuantitatif. Sample penelitian adalah lembar jawaban peserta didik kelas VII SMP Negeri yang berjumlah 119 lembar. Instrumen penelitian terdiri dari lembar soal ujian akhir semester, lembar jawaban siswa, dan rubrik penskoran. Data penelitian diperoleh dari lembar jawaban peserta didik yang dikumpulkan setelah ujian akhir semester mata pelajaran IPS berlangsung. Teknik analisis data menggunakan analisis statitik melalui program ITEMAN. Hasil penelitian menunjukkan nilai validitas dan reliabilitas instrumen dengan kategori tinggi. Hasil analisis tingkat kesukaran butir soal dan daya beda menunjukkan adanya ketidakseimbangan komposisi butir tes. Sehingga disimpulkan, disimpulkan bahwa soal tes ujian akhir semester mata pelajaran IPS SMP perlu dikaji ulang. Hasil penelitian ini diharapkan berimplikasi pada guru-guru dalam membuat soal tes mata pelajaran IPS kedepannya serta sebagai masukan bagi pemerintah dalam mengambil kebijakan khususnya dalam bidang pendidikan.

ABSTRACT

The analysis of the quality of the test items is very important because it will affect the results of the test itself. The problem is that teachers often do not analyze test items, so the test results do not accurately describe the potential of students. For this reason, this study aims to analyze the items on the semester final exam for social studies class VII. This type of research is a quantitative description. The research sample was the answer sheets of Class VII students from public middle schools, totaling 119 sheets. The research instrument consisted of end-of-semester exam question sheets, student answer sheets, and a scoring rubric. The research data was obtained from student answer sheets, which were collected after the semester final exams for social studies took place. The data analysis technique uses statistical analysis through the ITEMAN program. The results of the study place the validity and reliability of the instrument in the high category. The results of the analysis of the difficulty level of the items and the differential power showed that there was an imbalance in the composition of the test items. So it was concluded that the test questions for the end of semester exams for social studies subjects in junior high school needed to be reviewed. The results of this study are expected to have implications for teachers in making social studies test questions in the future as well as input for the government in making policies, especially in the field of education.

1. INTRODUCTION

Over the past few years, there has been a growing recognition of the significance of conducting research on item analysis to contribute to the enhancement of education quality, address the need for precise assessments that provide dependable outcomes, and elevate the standards of teaching and learning (Bichi, 2016; Gibta et al., 2020; Wijayanti, 2020). Consequently, the primary objective of this study will be to concentrate on the analysis of Social Studies final examination questions. Item analysis is important for refining the items that have been made by the teacher so that questions will be structured that have a function as a measuring tool for high-quality learning outcomes (Boopathiraj & Chellamani, 2013; Elgadal & Mariod, 2021). Some of the purposes of doing item analysis include (i) reviewing and analyzing the items to obtain quality questions, (ii) improving the quality of test items through the

*Corresponding author.
E-mail addresses: trizahra10019@gmail.com (Abdul Wahab)

revision of invalid questions, and (iii) obtaining items that are capable of diagnosing students' knowledge of the material being studied (Gao et al., 2020; Quaigrain & Arhin, 2017). A good item will be able to thoroughly and accurately diagnose students' abilities (Baran-Łucarz, 2019; Shirazi et al., 2019). So that the results of the tests carried out can correctly provide an overview of the potential of students. Conversely, bad items will have an impact on the results given, so that the teacher does not correctly know the potential of each student. This has an impact on the inaccuracy of the improvements made. So, it is not uncommon for new programs launched to not give maximum results in improving the quality of students. For this reason, educators need to know and pay attention to the quality of the items made before testing them on students. The quality of the good tests can be seen at first, validity is the degree of accuracy of a test in measuring what it is supposed to measure. The validity of each good item has a validity value above 0.6 (Taherdoost, 2016) . *Second,* reliability is the ability of a test to show the accuracy of measurement results. if validity is the determination of a test, reliability is the accuracy of a test (Taherdoost, 2016). T*hird,* the standard error of measurement (which is an estimate of how much error an evaluation researcher expects from a test that has been made). Fourth, item difficulty level; a good item is one with a medium difficulty level (Yunida & Arthur, 2023). This means that the questions tested are neither too easy nor too difficult. Ideally, in a test instrument, there are at least 25% of questions in the easy category, 50% of questions in the medium category, and 25% of questions in the difficult category. *Fifth,* discriminating power, namely the ability of a question to distinguish between students who are smart and students who are stupid, The distinguishing power of a minimum item of 0.40 in a good category.

But in reality, teachers often ignore the quality of the items on a test. Often, the original teacher makes questions about the material that has been studied and tests them directly on students without considering the composition of the questions in a test. This happened in social studies subjects at public junior high schools, where the teacher did not care about the quality of the final semester test items. Findings in the initial observation of researchers indicated that the social studies teacher at the public junior high school stated that the questions for the end of semester exams were questions sent from the agency. The teacher directly tests these questions on students and checks student answer sheets according to the answer keys from the department without doing item analysis. The teacher did not analyze the questions because they had limited time. Several studies also show that teachers have never analyzed the test items given to students. In fact, teachers at school more often pick questions from books than make up their own questions. Even though the questions contained in the book are not necessarily in accordance with the learning objectives that have been implemented, as a result, many questions are not appropriate and even deviate from the learning that has been done. The results of previous research analysis of quality control test items, both qualitatively and quantitatively, have never been carried out in the city of Yogyakarta, so from year to year, the quality of quality control test items is still unknown (Wibawa, 2019). The results of other study show that the math questions that have been tested on students of SMA Negeri 1 Purbalingga have never been analyzed before, both qualitatively and quantitatively, since the change of the educational curriculum from the education level unit curriculum (KTSP) to the 2013 curriculum (Suzana, 2018). Therefore, the validity, reliability, level of difficulty, and discriminating power of these questions are unknown.

Based on the problems above, research on the analysis of urgent items was carried out to produce high-quality tests. The quality of the test will have an impact on the process and results of student assessment. Assessment is an important aspect of the learning process because it is a process to determine student competency achievement during and after participating in the learning process (Ningsih et al., 2019; Zhampeissova et al., 2020). Assessment is carried out in an integrated manner to reveal all aspects of students' abilities both in terms of knowledge, skills, and attitudes (Cahyadi et al., 2022; Kolovou et al., 2021). Assessment serves three purposes: (i) identifying students' strengths and weaknesses so that their weaknesses can be improved; determining the success of a program's implementation so that, if it is still insufficient, adjustments can be made to methods, approaches, strategies, and learning techniques; and (iii) serving as a benchmark for future performance (Black & Wiliam, 2018; Hairida & Junanto, 2018; Leber et al., 2018). Based on the goal and function of the assessment, it is understood that assessment activities are important in the educational process in order to create a better education. Based on the aforementioned issues, the study intends to use ITEMAN to analyze the final semester exam questions for social studies subjects in Public Middle Schools. The ITEMAN application is a computer program commonly used to analyze multiple-choice questions. ITEMAN is an empirical item analysis with a classic approach model used to determine the quality of an item or a test (Aulia et al., 2014; Permatasari et al., 2019). The results of the analysis of the items using ITEMAN include the level of validity of the items, instrument reliability, level of difficulty, and differential power of the items. The advantages of ITEMAN are that: (1) in inputting data, it can be assisted by the Microsoft Excel program; (2) the results of the analysis are easy to understand because they are in the

form of; (3) the appearance of ITEMAN is attractive; and (4) it has many versions (Clauser & Hambleton, 2017; Suwarto, 2021). Several previous relevant studies have examined similar topics. For instance study that conducted using the ITEMAN program to assess English test items in an online learning setting (Ma'rifah et al., 2021). In comparison, our study applies the ITEMAN program to analyze Social Studies examination items in junior high schools. Another study also utilized the ITEMAN program but focused on evaluating the difficulty level, validity, and reliability of test items specifically designed for visually impaired children (Shakir, 2021). In contrast, our study concentrates on utilizing the ITEMAN program to analyze the items within the end-of-semester examination for Social Studies. Additionally, other study employed the ITEMAN program to analyze general examination questions for junior high school (Suwarto, 2021). In contrast, this study narrows its focus to the use of the ITEMAN program for analyzing Social Studies examination items.

By analyzing previous relevant studies, it can be concluded that there are several gaps that highlight the significance of the research on Item Analysis of Final Examination Questions for Social Studies in Junior High Schools through the ITEMAN Program. These gaps include the following: (i) previous studies utilizing ITEMAN generally lacked focus on Social Studies in junior high schools. Therefore, conducting this research is essential to address this gap and provide more precise insights into the analysis of Social Studies examination questions in junior high schools using the ITEMAN Program. (ii) While previous studies using ITEMAN encompassed various types of analyses, the specific analysis of end-of-semester examination questions was rarely explored. Consequently, this research holds importance as it will contribute to a deeper understanding of the characteristics and quality of Social Studies examination questions at the end of the semester through the utilization of the ITEMAN Program. Considering these gaps, this research plays a critical role in complementing prior studies, offering a more specific understanding of the analysis of Social Studies examination questions in junior high schools, and supporting the enhancement of the quality of Social Studies education. As a result, the purpose of this study is to analyze the semester test questions provided to junior high school students learning social studies using the ITEMAN programs.

2. METHODS

This research uses a type of quantitative descriptive with the cross-sectional study design. The reserach aims to describe a situation objectively by using numbers, starting from data collection, data interpretation, and drawing a conclusion. The purpose of this description is to help the reader know what is happening in the environment under observation and what kinds of events or activities occur in the research setting. The cross-sectional study design was used because the research was conducted at a certain point in time, namely at the time of data collection for the social studies final exam questions (Wijaya, 2013). The research design was able to provide a clear picture of the quality of social studies final exam questions at the junior high school level at that time. The description of this research design is show in Figure 1. The research subjects were junior high school students. The population in this study was all state junior high school students. The research sample consisted of class VII students who took the semester final exams on social studies learning, totaling 119 students who were taken by the technique of purposive sampling. The research utilized various instruments, such as question sheets for the end-of-semester exams in the seventh-grade social studies class, along with the accompanying answer sheets and answer keys provided by the school. To collect data, a document study was conducted, wherein the researchers analyzed the student answer sheets subsequent to the completion of the final semester exams for the social studies subject.

Data collection techniques in this study were carried out through interview techniques and documentation techniques. Interview techniques were conducted to collect initial data about the problems to be studied. Then the documentation technique is carried out to obtain a grid of writing items, exam questions, and student answer sheets. The item grid for the final school exam for social studies. There are 50 questions, but only 40 indicators. In an ideal test scenario, the number of questions should be equal to the number of indicators. However, in this case, there is a discrepancy between the number of questions and indicators, with more questions than indicators. This imbalance can result in an uneven coverage of the material being tested. Analysis techniques were carried out through statistical analysis with the help of the ITEMAN program. Tests of validity, reliability, item difficulty level, and discriminatory power were involved in analyzing the final semester exam questions for social studies subjects. Test the validity of the items to see if they are suitable for the ability being measured. The validity of the question items follows the following criteria as show in Table 1.

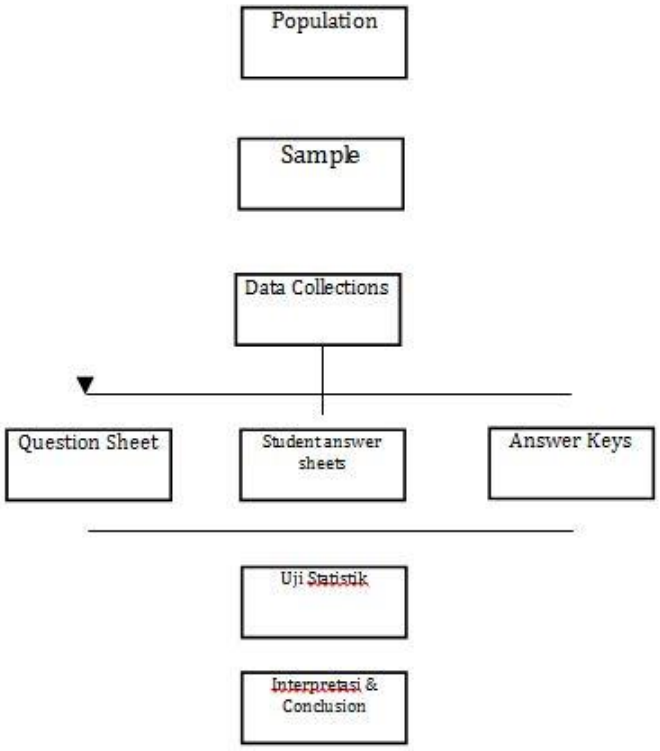


Figure 1. Design Research of the cross-sectional study

Table 1. Validity Criteria

Koefisien Korelasi	Category
0.81-1.00	Very high correlation
0.61-0.80	High correlation
0.41-0.60	Enough correlation
0.21-0.40	Low correlation
0.00-0.20	Very low correlation

Furthermore, the reliability test was carried out to determine the level of accuracy of an instrument in measuring what should be measured. The results of the reliability test follow the following criteria as show in Table 2.

Table 2. Reliability Criteria

Reliability Index	Category
0.81-1.00	Very high
0.61-0.80	High
0.41-0.60	Enough
0.21-0.40	Low
<0.20	Very Low

Base on Table 2 an instrument is said to be reliable when the minimum alpha value is 0.7. In addition to the analysis of validity and reliability, the analysis of the items also analyzed the level of difficulty and discriminating power. The difficulty level of the items and the discriminating power follows the following criteria as show in Table 3.

Table 3. The Difficulty Level Category

Score of Prop. Correct (p)	Category
$0 < P \leq 0,3$	Difficult
$0,3 < P \leq 0,7$	Middle
$0,3 < P \leq 0,7$	Easy

Base on Table 3 it is know that question details are said to be good when they are in the medium category. The criteria of point biserial is show in Table 4.

Table 4. Criteria of Point Biserial

Score of Point Biserial (D)	Category
-1 < D ≤ 0,19	Not Good
0,2 < D ≤ 0,2.9	Low (Revision Require)
0,3 < D ≤ 0,3.9	Middle (There is no need to revise)
0,4 < D ≤ 1	Good

Base on Table 4 differential power of at least 0.4. Good question. The higher the discriminating power, the better the test item performs.

3. RESULT AND DISCUSSION

Results

Validitas Item

We tested 50 items on the social studies final exam in the form of multiple choices. The results of the analysis are presented as show in Table 5.

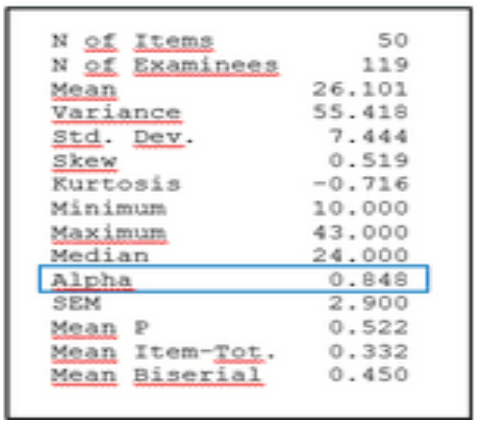
Table 5. Score of Validitas Item

Number of Question Items	Coefisien of correlation
14	0.81-1.00
26	0.61-0.80
10	0.41-0.60
0	0.21-0.40
0	0.00-0.20
Mean	**0.72**

Based on Table 5, it is known that in general, all items of multiple-choice questions have a validity value of 0.72 and fall into the high correlation category. This means that all items, on average, can measure what should be measured. However, when translated, there are 10 items that are in the category of moderate correlation. These 10 questions still need to be reviewed again. According to the research findings, each question indicator should ideally represent each question. It turned out that, based on the analysis of the semester final exam questions for social studies subjects, there were 40 indicator questions with 50 questions. This means that there is one indicator with two questions. So, this needs to be improved for future exam questions.

Reliabilitas

The results of the instrument reliability test of 50 items on the Social Sciences subject in junior high school are presented in Figure 1.



N of Items	50
N of Examinees	119
Mean	26.101
Variance	55.418
Std. Dev.	7.444
Skew	0.519
Kurtosis	-0.716
Minimum	10.000
Maximum	43.000
Median	24.000
Alpha	0.848
SEM	2.900
Mean P	0.522
Mean Item-Tot.	0.332
Mean Biserial	0.450

Figure 1. Score of Reliabilitas Instrument

Base on Figure 1 the reliability of the instrument, as shown by the alpha value of 0.848 in the very high category, is known from the figure above. This indicates that the IPS Grade VII final exam questions are highly reliable in assessing student potential. As a result, they are highly effective.

The items' difficulty level

The results of the analysis of the difficulty level of the items are presented as show in Table 6.

Table 6. The results of the Analysis of the Difficulty Level

Prop. Correct (p)	Category	Question number	Number of questions	Persentase
0 < P ≤ 0,3	Difficult	5, 8, 9, 16, 17,26, 27, 31, 33, 38	10 Questions	20%
0.3 < P ≤ 0,7	Middle	1, 4, 6, 7, 11, 13, 15, 19,20,22, 23, 24, 25, 29, 32, 34, 36, 37, 40, 43, 44, 45, 48, 49, 50	25 Questions	50%
0.3 < P ≤ 0,7	Easy	2, 3, 10, 12, 14, 18, 21, 28, 30, 35, 39, 41, 42, 46, 47	15 Questions	30%

Based on Table 6, it is known that there are 25 questions in the medium category, 15 questions in the easy category, and 10 questions in the difficult category. Based on the table above, the composition of the questions is 20% difficult questions, 50% medium questions, and 30% easy questions. When compared with the ideal percentage, which is 25% easy and difficult questions and 50% medium questions, then the final semester exam questions for IPS subjects for the easy and difficult categories still need to be reviewed because they do not match the ideal indicators of questions that are said to be good.

Point Biserial

The results of the analysis of point biserial for 50 multiple choice items in the semester final exam for social studies class VII are presented in Table 7.

Table 7. The results of the Analysis of Point Biserial

Point Biserial (D)	Category	Question number	Number of questions
-1 < D ≤ 0,19	Not Good	-	-
0,2 < D ≤ 0,2.9	Low (Revision Require)	2, 9, 11, 14, 17, 27, 29, 31, 32, 33, 35, 38, 39, 44, 46, 47, 49, 50	18
0,3 < D ≤ 0,3.9	Middle (There is no need to revise)	1, 3, 6, 10, 12, 18, 20, 21, 22, 24, 25, 28, 30, 37, 40, 42, 48	17
0,4 < D ≤ 1	Good	4, 5, 7, 8, 13, 15, 19, 23, 25, 26,, 34, 36, 41, 43, 45	15

Based on Table 7 it can be concluded that there are 18 items in the low category, which means that the questions are not able to distinguish between the upper group (smart students) and the lower group (less intelligent students). There are 17 items in the moderate category and 15 items in the good category. For the medium category, it means that the questions can be answered by the upper-group and also the lower-group students. For Different Power, a high category means that the questions can only be answered by students from the upper group, and students from the lower group have difficulty answering these questions. Based on the results in the table above, it is concluded that of the 18 items that have different power in the low category, it is necessary to review the items in order to improve the quality of the questions on the next exam.

Discussion

The results of the analysis of 50 multiple-choice items on the semester final exam questions for social studies class VII in public middle schools show that the items are valid and reliable, with high validity values and very high reliability values. However, there are 10 items that need to be reviewed because, based on the validity of the item, these 10 items are still in the sufficient category. Based on the identification results, the number of question indicators and the number of item items are not the same. There are 40 question indicators with a total of 50 items. This means that there is 1 indicator question

with 2 questions. This is a consideration for the teacher to review the item questions, because ideally, 1 item indicator represents 1 item. When creating an accurate and reliable exam, the disparity between the amount of indicator questions and test items might be problematic (Clark & Watson, 2019; Reynolds et al., 2021). The mismatch between the number of indicators and the number of test items might lead to items being created unevenly when testing each indicator (Ivars-Baidal et al., 2021; Mai et al., 2021; Surucu & Maslakci, 2020). This may result in unreliable test results and a cloudy image of the ability being assessed. The results of this study support several studies that show that good questions are those with the same number of question indicators as the number of items (Kim et al., 2019; Qurrota et al., 2022; Schaeffer & Dykema, 2020). Because in order to obtain accurate and comprehensive student-potential diagnostic results, each indicator question must produce one item. To make the right question indicators, the teacher must pay attention to the material to be tested, learning indicators, basic competencies, and competency standards (Prasetyono et al., 2021; Rintayati et al., 2020; Said & Muslimah, 2021).

For the level of difficulty of the items, based on the findings of the study, it was concluded that the composition of the items in the semester final exams for social studies subjects in junior high school did not meet the ideal criteria. The results of the analysis show that the number of easy questions exceeds the ideal conditions, and the number of difficult questions is less than the ideal conditions. Thus, it is necessary to review the two categories of questions. The results of this study support several studies that the exam could be negatively impacted because of the disparity in the number of item difficulty ratios (Gazi et al., 2022; Herbert et al., 2020; Özer & Bilgisi Öz, 2020). Test results between participants cannot be reliably compared if the ratio of item difficulty levels does not match (Din, 2020; Rozental et al., 2019). This is due to the possibility that some participants may struggle to respond to the simpler questions while others may be able to respond to the more challenging ones. In order to create a good exam, it is crucial for teachers to pay attention to the ratio of item difficulty levels.

Several studies confirm that good questions are those that have a percentage of difficult questions as much as 25% of the total number of test items, easy questions as much as 25% of the total number of test items, and moderate questions as much as 50% of the total number of test items If the composition of the items in an exam text is unbalanced, then the ability information generated will also not be normally distributed (Dewi et al., 2019; Warju et al., 2020; Weller et al., 2018). Even so, there are those who argue that the questions that are considered good are moderate questions, that is, questions whose index difficulty ranges from 0.26 to 0.75 (Quaigrain & Arhin, 2017). The various criteria have the tendency to say that the items with a difficulty index of less than 0.25 and more than 0.75 should be avoided or not used because they are too difficult or too easy and are therefore less reflective as good measuring tools (Adiga et al., 2021; Ibrahim & Yahia, 2021). The questions that are difficult need to be included in a test to motivate students who are smart, while the questions that are easy will raise the spirits of weak students. A balanced level of difficulty between difficult, medium, and easy questions will be able to distinguish students who are less intelligent from those who are good at it.

It is known that the question only asks for the definition of space at the thinking level of C1. The answer options given were very easy for students to guess; the distractor didn't work. So, that the question is easily answered by students. A good question is one where the answer is correct and the distractor works well. Most of the correct answer keys must be answered correctly by students in the upper group, while the distractor is mostly chosen by the lower group. In addition, the length of the answer options must also be relatively the same. As can be seen from the questions above, the statement of option C is very different from the other answer options. so that the question is categorized as easy. Some of the factors that cause the above questions to fall into the "difficult" category include (i) using a stimulus in the form of an image that must be analyzed by students. This stimulus is actually able to stimulate students' high-order thinking skills, but for students who have moderate or even low abilities, a stimulus like the picture above becomes difficult for them (Doley, 2023; Moses et al., 2019); (ii) the length of the statement in the answer choices is relatively the same, so that the answer key is not easy for students to guess; and (iii) the questions above ask about concepts, so students must really understand and know the difference between each concept in the answer choices presented (Zohar & Alboher Agmon, 2018). Based on the two examples of social studies class VII final exam questions presented above, it can be seen why these questions fall into the easy or difficult categories. so that the teacher can take action on questions that fall into the easy or difficult category. When the preparation of the questions meets the criteria, the quality of the items will also increase, so that a question is not only a measuring tool that does not measure anything to be measured, but each domain of students or things that must be evaluated by students can be illustrated from the questions that have been prepared. An organized test based on the principles and procedures for preparing tests will produce tests of good quality. This is in line with what was expressed by study who suggested that the test should be prepared according to the principles and procedures for preparing tests (Wibowo & Faizah, 2021). Therefore, teachers should be able to improve

the quality of the tests they compose so that the tests they give to students are of good quality. Analysis of the items can also help the teacher improve the quality of the items that have been answered. The item analysis is designed to find defects in test items so that they can be corrected before being used in the next test, as well as to find out if the test given is too difficult or too easy for students to do (Gibta et al., 2020; Nengsi & Efrina, 2019). An analysis of the items needs to be done to determine the extent to which the items can be used in testing and as a control for student achievement results. Based on the research findings, it appears that there are 18 items with low discriminating power. This means that the questions were not able to distinguish the abilities of participants in the upper and lower groups. So that the 18 questions with low differential power need to be reviewed again in order to obtain quality test questions. Discriminating power is the ability of the items to distinguish participants who are able and those who are less able. The test is said to have good discriminating power if it is given to students who have high abilities; the results are good; but if it is given to students who have low abilities, the students have low scores as well. Low test ability in differentiating between students with high and low abilities might result from the disparity between the differential power of the test items and the ideal circumstances (Jin et al., 2020; Mrkva et al., 2021). The test results won't be able to accurately depict the test taker's skills if the items can't tell the difference between test takers with high and low competency levels. Several studies have shown that a good test is one with high discriminating power (Perdana et al., 2019; Sarwanto et al., 2020). This is because the higher the coefficient of discriminating power of an item, the more capable the item is of distinguishing between students who master competence and those who lack competence (Pittman et al., 2020; Utama et al., 2020). The test is said to have no discriminating power, and especially if it is tested on weak children, the results are higher or if given to both categories of students, the results are the same. Thus, tests that do not have discriminatory power will not produce results that are in accordance with the actual abilities of students. It's really strange if a smart child doesn't pass, but a stupid child passes well without being manipulated by the assessor or by coincidence.

In addition, in terms of the level of difficulty of the items and the differential power, it also indicates the need for revision of the exam questions. The questions for the end of semester exams for social studies subjects in state junior high schools are considered disproportionate because they are not balanced between difficult questions, moderate questions, and easy questions. Likewise, with discriminating power, it shows that there are more questions that have low discriminating power, which means that the items are less able to distinguish students with high abilities from students with low abilities. So it is necessary to do a reassessment. The researcher recommends that each teacher conduct an analysis of the test items given to students to determine the quality of the questions and obtain higher-quality questions. Research on the analysis of these items was limited to final semester exam questions for social studies subjects at public junior high schools. So that further research with a wider sample is needed to determine the quality of exam questions in a more comprehensive manner in schools.

4. CONCLUSION

The findings of the study indicate that the social studies questions at the end of the semester in public middle schools need to be re-examined or revised. This is obtained from the validity value of the item, which indicates a discrepancy between the question indicators and the item items. It was concluded that the semester final exam questions for social studies in public middle schools need to be reviewed so that they are in accordance with the characteristics of good item items. It is intended that the tests given are able to accurately diagnose the real abilities of students. As a result, the results provided can serve as a benchmark for future improvements for both teachers and the government.

5. REFERENCES

Adiga, M. N. S., Acharya, S., & Holla, R. (2021). Item Analysis of Multiple-Choice Questions in Pharmacology in an Indian Medical School. *Journal of Health and Allied Sciences NU*, *11*(3), 130–135. https://doi.org/10.1055/s-0041-1722822.

Aulia, I. F., Sukirlan, M., & Sudirman. (2014). Analysis of the Quality of Teacher-Made Reading Comprehension Test Items Using ITEMAN. *Unila Journal of English Teaching*, *3*(4). https://www.neliti.com/publications/194356/analysis-of-the-quality-of-teacher-made-reading-comprehension-test-items-using-i.

Baran-Łucarz, M. (2019). Formative assessment in the English as a foreign language classroom in secondary schools in Poland. Report on a mixed-method study. *Journal of Education Culture and Society*, *10*(2), 309–327. https://doi.org/10.15503/jecs20192.309.327.

Bichi, A. A. (2016). Classical test theory : An introduction to linear modeling approach to test and item

analysis. *International Journal for Social Studies*, *2*(9). https://edupediapublications.org/journals

Black, P., & Wiliam, D. (2018). Classroom Assessment and Pedagogy. *Assessment in Education: Principles, Policy and Practice*, 1–25. https://doi.org/https://doi.org/10.1080/0969594X.2018.1441807.

Boopathiraj, C., & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. *International Journal of Social Science & Interdisciplinary Research*, *2*(2). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a9e08c03848e95760275e36f75cae88e49bc6c65.

Cahyadi, W., Aswita, D., & Ningsih, T. Z. (2022). Analysis of The Development of Non-Cognitive Assessment Instrument to Support Online History Learning in Jambi City High School. *AL-ISHLAH: Jurnal Pendidikan*, *14*(3), 3265–3274. https://doi.org/10.35445/alishlah.v14i3.2044.

Clark, L. A., & Watson, D. (2019). Constructing Validity: New Developments in Creating Objective Measuring Instruments. *Psychological Assessment*, *176*(3), 1412. https://doi.org/10.1037/pas0000626.Constructing.

Clauser, J. C., & Hambleton, R. K. (2017). Item Analysis for Classroom Assessments in Higher Education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 355–369). Routledge.

Dewi, S. S., Hariastuti, R. M., & Utami, A. U. (2019). Analisis Tingkat Kesukaran Dan Daya Pembeda Soal Olimpiade Matematika (Omi) Tingkat Smp Tahun 2018. *Transformasi : Jurnal Pendidikan Matematika Dan Matematika*, *3*(1), 15–26. https://doi.org/10.36526/tr.v3i1.388.

Din, M. (2020). Evaluating university students' critical thinking ability as reflected in their critical reading skill: A study at bachelor level in Pakistan. *Thinking Skills and Creativity*, *35*(September 2019), 100627. https://doi.org/10.1016/j.tsc.2020.100627.

Doley, S. K. (2023). Stimulus Appraisal-Based L2 Attitude and Motivation among Indian ESL Learners. *International Journal of Instruction*, *16*(2), 603–622. https://doi.org/10.29333/iji.2023.16232a.

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*, *16*(3), 334–346. https://doi.org/10.18502/sjms.v16i3.9695.

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education*, *7*(1), 1–14. https://doi.org/10.1186/s40594-020-00225-4.

Gazi, F., Atan, T., & Kılıç, M. (2022). The Assessment of Internal Indicators on The Balanced Scorecard Measures of Sustainability. *Sustainability (Switzerland)*, *14*(14), 1–19. https://doi.org/10.3390/su14148595.

Gibta, Melani, M., Susanti, I., & Dharma, U. S. (2020). Item Analysis of Force Material Problem in Elementary School. *Jurnal Pendidikan Sekolah Dasar*, *3*(April), 23–32. https://repository.usd.ac.id/37825/1/6315_31431-89165-1-PB.pdf.

Hairida, H., & Junanto, T. (2018). The Effectiveness of Performance Assessment in Project-Based Learning by Utilizing Local Potential to Increase the Science Literacy. *International Journal of Pedagogy and Teacher Education*, *2*(July), 17. https://doi.org/10.20961/ijpte.v2i0.25722.

Herbert, J. S., Mitchell, A., Brentnall, S. J., & Bird, A. L. (2020). Identifying Rewards Over Difficulties Buffers the Impact of Time in COVID-19 Lockdown for Parents in Australia. *Frontiers in Psychology*, *11*(December), 1–11. https://doi.org/10.3389/fpsyg.2020.606507.

Ibrahim, A., & Yahia, O. (2021). Post-validation item analysis to assess the validity and reliability of multiple- choice questions at a medical college with an innovative curriculum. *Medical Education*, *34*(6), 359–362. https://www.researchgate.net/profile/Amar-Yahia-2/publication/361839573_Post-validation_item_analysis_to_assess_the_validity_and_reliability_of_multiple-choice_questions_at_a_medical_college_with_an_innovative_curriculum/links/62de9f57aa5823729ee0bce6/Post-validation-item-analysis-to-assess-the-validity-and-reliability-of-multiple-choice-questions-at-a-medical-college-with-an-innovative-curriculum.pdf.

Ivars-Baidal, J. A., Celdrán-Bernabeu, M. A., Femenia-Serra, F., Perles-Ribes, J. F., & Giner-Sánchez, D. (2021). Measuring the progress of smart destinations: The use of indicators as a management tool. *Journal of Destination Marketing and Management*, *19*, 100531. https://doi.org/10.1016/j.jdmm.2020.100531.

Jin, K. Y., Reichert, F., Cagasan, L. P., de la Torre, J., & Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring test dimensionality and performance differences. *Computers and Education*, *157*(June), 103968. https://doi.org/10.1016/j.compedu.2020.103968.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of

Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406.

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A. K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education*, *100*(April), 103298. https://doi.org/10.1016/j.tate.2021.103298.

Leber, J., Renkl, A., Nückles, M., & Wäschle, K. (2018). When the type of assessment counteracts teaching for understanding. *Learning: Research and Practice*, *4*(2), 161–179. https://doi.org/10.1080/23735082.2017.1285422.

Ma'rifah, U., Algiovan, N., & Sutarsyah, C. (2021). An Item Analysis of English Test During Online Learning. *International Journal of Multicultural and Multireligious Understanding*, *8*(12), 647–654. https://doi.org/10.18415/ijmmu.v8i12.3396.

Mai, R., Niemand, T., & Kraus, S. (2021). A tailored-fit model evaluation strategy for better decisions about structural equation models. *Technological Forecasting and Social Change*, *173*(August), 121142. https://doi.org/10.1016/j.techfore.2021.121142.

Moses, D. A., Leonard, M. K., Makin, J. G., & Chang, E. F. (2019). Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-10994-4.

Mrkva, K., Posner, N. A., Reeck, C., & Johnson, E. J. (2021). Do Nudges Reduce Disparities? Choice Architecture Compensates for Low Consumer Knowledge. *Journal of Marketing*, *85*(4), 67–84. https://doi.org/10.1177/0022242921993186.

Nengsi, A. R., & Efrina, G. (2019). Optimalisasi Tes Prestasi Buatan Guru Mata Pelajaran Ips Sd Untuk Evaluasi Pembelajaran Yang Presisi. *Ta'dib*, *22*(2), 121. https://doi.org/10.31958/jt.v22i2.1498.

Ningsih, T. Z., Sariyatun, & Sutimin, L. A. (2019). Development of portfolio assessment to measure the student's skill of using primary source evidence. *New Educational Review*, *56*(2), 101–113. https://doi.org/10.15804/tner.2019.56.2.08.

Özer, M., & Bilgisi Öz, M. (2020). What Does PISA Tell Us About Performance of Education Systems? PISA Eğitim Sistemlerinin Performansı Hakkında Bize Ne Söylüyor? *Bartın University Journal of Faculty of Education*, *9*(2), 217–228. https://doi.org/10.14686/buefad.697153.

Perdana, R., Riwayani, R., Jumadi, J., & Rosana, D. (2019). Development, Reliability, and Validity of Open-ended Testto Measure Student's Digital Literacy Skil. *International Journal of Educational Research Review*, *4*(4), 504–516. https://doi.org/10.24331/ijere.628309.

Permatasari, B. D., Gunarhadi, & Riyadi. (2019). The influence of problem based learning towards social science learning outcomes viewed from learning interest. *International Journal of Evaluation and Research in Education*, *8*(1), 39–46. https://doi.org/10.11591/ijere.v8i1.15594.

Pittman, R. T., Zhang, S., Binks-Cantrell, E., Hudson, A., & Joshi, R. M. (2020). Teachers' knowledge about language constructs related to literacy skills and student achievement in low socio-economic status schools. *Dyslexia*, *26*(2), 200–219. https://doi.org/10.1002/dys.1628.

Prasetyono, H., Abdillah, A., Djuhartono, T., Ramdayana, I. P., & Desnaranti, L. (2021). Improvement of teacher's professional competency in strengthening learning methods to maximize curriculum implementation. *International Journal of Evaluation and Research in Education*, *10*(2), 720–727. https://doi.org/10.11591/ijere.v10i2.21010.

Quaigrain, K., & Arhin, A. K. (2017). Using Reliability and Item Analysis to Evaluate A Teacher-Developed Test in Educational Measurement and Evaluation. *Cogent Education*, *12*(1). https://doi.org/10.1080/2331186X.2017.1301013.

Qurrota, A. A. S., Siskawati, F. S., & Irawati, T. N. (2022). Analisis Kelayakan Butir Soal pada Media INTERMATHLY (Interesting Mathematic Monopoly). *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *6*(1), 634–654. https://doi.org/10.31004/cendekia.v6i1.1181.

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). The Problem of Bias in Psychological Assessment. In C. R. Reynolds, R. A. Altmann, & D. N. Allen (Eds.), *Mastering Modern Psycological Testing* (pp. 573–613). Springer US.

Rintayati, P., Lukitasari, H., & Syawaludin, A. (2020). Development of Two-Tier Multiple Choice Test to Assess Indonesian Elementary Students' Higher-Order Thinking Skills. *International Journal of Instruction*, *14*(1), 555–566. https://doi.org/10.29333/IJI.2021.14133A.

Rozental, A., Kottorp, A., Forsström, D., Månsson, K., Boettcher, J., Andersson, G., Furmark, T., & Carlbring, P. (2019). The Negative Effects Questionnaire: Psychometric properties of an instrument for assessing negative effects in psychological treatments. *Behavioural and Cognitive Psychotherapy*, *47*(5), 559–572. https://doi.org/10.1017/S1352465819000018.

Said, A., & Muslimah, M. (2021). Evaluation of Learning Outcomes of Moral Faith Subjects during Covid-19 Pandemic at MIN East Kotawaringin. *Bulletin of Science Education*, *1*(1), 13–26.

https://doi.org/10.51278/bse.v1i1.99.
Sarwanto, Fajari, L. E. W., & Chumdari. (2020). Open-Ended Questions to Assess Critical-Thinking Skills in Indonesian Elementary School. *International Journal of Instruction*, *14*(1), 615–630. https://doi.org/10.29333/IJI.2021.14137A.
Schaeffer, N. C., & Dykema, J. (2020). Advances in the Science of Asking Questions. *Annual Review of Sociology*, *46*(March), 37–60. https://doi.org/10.1146/annurev-soc-121919-054544.
Shakir, M. A. (2021). Assessment of Learning Achievement of Visually Impaired Children at Primary Level. *Pakistan Journal of Educational Research and Evaluation*, *9*(2), 44–52. http://journals.pu.edu.pk/journals/index.php/PJERE/article/view/5311.
Shirazi, M. A., Alavi, S. M., & Salarian, H. (2019). An Investigation into Item Types and Text Types of Reading Comprehension Section of Iranian Ph.D. Entrance Exams Using G-theory. *Journal of Modern Research in English Language Studies*, *6*(1), 1–29. https://doi.org/10.30479/jmrels.2019.10591.1326.
Surucu, L., & Maslakci, A. (2020). Business & Management Studies : *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540.
Suwarto, S. (2021). The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students Endonezya İ kinci Yar ı y ı l Sekizinci S ı n ı f Ö ğ rencileri İ çin Final S ı nav ı n ı n Özellikleri. *Turkish Online Journal of Qualitative Inquiry (TOJQI)*, *12*(9), 356–370. https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=13096591&AN=160604451&h=K2DbfvLAoTNBnejP68bkGfn5XmFXzQcBX55uayj0k6Qa8zBEUzB8aJR83OL1kTIwX0GlzBBmysM4r1uFMZ1dGA%3D%3D&crl=c.
Suzana, A. (2018). Analisis Tingkat Kesukaran dan Daya Beda Butir-Butir Soal Penilaian Akhir Tahun Matematika Kelas X di SMA Negeri 1 Purbalingga. *MathGram Matematika*, *2*(2), 1–8. https://ejournal.unugha.ac.id/index.php/mthg/article/view/172.
Taherdoost, H. (2016). Validity and Reliability of the Research Instrument ; How to Test the Validation of a Questionnaire / Survey in a Research Hamed Taherdoost To cite this version : HAL Id : hal-02546799 Validity and Reliability of the Research Instrument ; How to Test the. *International Journal of Academic Research in Management*, *5*(3), 28–36. https://doi.org/10.2139/ssrn.3205040.
Utama, C., Sajidan, Nurkamto, J., & Wiranto. (2020). The instrument development to measure higher-order thinking skills for pre-service biology teacher. *International Journal of Instruction*, *13*(4), 833–848. https://doi.org/10.29333/iji.2020.13451a.
Warju, W., Ariyanto, S. R., Soeryanto, S., & Trisna, R. A. (2020). Analisis Kualitas Butir Soal Tipe Hots Pada Kompetensi Sistem Rem Di Sekolah Menengah Kejuruan. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *17*(1), 95. https://doi.org/10.23887/jptk-undiksha.v17i1.22914.
Weller, S. C., Vickers, B., Russell Bernard, H., Blackburn, A. M., Borgatti, S., Gravlee, C. C., & Johnson, J. C. (2018). Open-ended interview questions and saturation. *PLoS ONE*, *13*(6), 1–18. https://doi.org/10.1371/journal.pone.0198606.
Wibawa, E. A. (2019). Karakteristik Butir Soal Tes Ujian Akhir Semester Hukum Bisnis. *Jurnal Pendidikan Akuntansi Indonesia*, *17*(1), 86–96. https://doi.org/10.21831/jpai.v17i1.26339.
Wibowo, T. E., & Faizah, S. (2021). Pengembangan Soal Tes Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Pada Materi Bentuk Aljabar. *Alifmatika: Jurnal Pendidikan Dan Pembelajaran Matematika*, *3*(2), 145–158. https://doi.org/10.35316/alifmatika.2021.v3i2.145-158.
Wijaya, I. M. K. (2013). Pengetahuan, sikap dan motivasi terhadap keaktifan kader Dalam pengendalian tuberkulosis. *Jurnal Kesehatan Masyarakat*, *8*(2), 137–144. https://doi.org/10.15294/kemas.v8i2.2637.
Wijayanti, P. S. (2020). Item Quality Analysis For Measuring Mathematical Problem-Solving Skills. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *9*(4), 1223–1234. https://doi.org/10.24127/ajpm.v9i4.3036.
Yunida, H., & Arthur, R. (2023). Bloom ' s Taxonomy Approach to Cognitive Space Using Classic Test Theory and Modern Theory. *East Asian Journal of Multidisciplinary Research (EAJMR)*, *2*(1), 95–108. https://doi.org/10.55927/eajmr.v2i1.2331.
Zhampeissova, K., Alena, G., Ekaterina, V., & Zhanna, E. (2020). "Academic Performance and Cognitive Load in Mobile Learning." *International Journal of Interactive Mobile Technologies*, *14*(21), 78–91,. https://doi.org/10.3991/ijim.v14i21.18439.
Zohar, A., & Alboher Agmon, V. (2018). Raising test scores vs. teaching higher order thinking (HOT): senior science teachers' views on how several concurrent policies affect classroom practices. *Research in Science and Technological Education*, *36*(2), 243–260. https://doi.org/10.1080/02635143.2017.1395332.