Journal of Educational Sciences
Journal homepage: https://jes.ejournal.unri.ac.id/index.php/JES

P-ISSN 2581-1657
E-ISSN 2581-2203

Evaluation of Item Quality: Analysis of Difficulty Level and Distinction Power with Quantitative Methods

Annora Pratama Putri*, Joko Sayono
Master of History Education, State University of Malang, Malang, 65145 , Indonesia

ARTICLE INFO

Article history:

Received: 27 Nov 2025
Revised: 13 Des 2025
Accepted: 24 Des 2025
Published online: 05 Jan 2026

Keywords:

Item Quality, Difficulty, Test

* *Corresponding author:*

E-mail: anorapratamaputri12@gmail.com

Article Doi:

https://doi.org/10.31258/jes.10.1.p.317-330

A B S T R A C T

Evaluation is a systematic process of collecting and analyzing information to assess the quality of a program or product. This study aims to evaluate the quality of multiple-choice questions on the Japanese Occupation and Indonesian Independence topic through analysis of difficulty level and discriminating power. The main problem in this study was the lack of an evaluation instrument capable of accurately measuring learning outcomes and proportionally differentiating student abilities. The method used was quantitative, with sample answer sheets of 11th-grade students who completed 20 multiple-choice questions and 5 essay questions. Difficulty level analysis was conducted to determine the extent to which students could answer the questions, while discriminatory power was used to assess the ability of the questions to differentiate between high-ability and low-ability students. The results showed that 90% of the multiple-choice questions were classified as easy, while the essay questions had a more even distribution of difficulty levels. In terms of discriminatory power, some questions showed a fair to good category, but some had low or even negative discrimination power. These findings emphasize the need for revision of several questions to improve the validity and effectiveness of the instrument. Periodic evaluation of question quality is crucial to ensure assessments that support the quality of history learning.

1. Introduction

Evaluation refers to the process of determining the extent to which learning objectives have been achieved and involves assessing both the outcomes and the learning process. Based on recent research, evaluation can be divided into several categories, such as formative and summative evaluation, each of which has different purposes and applications in educational contexts (Yuwono & Mirnawati, 2021). In the history learning process at the high school level, evaluation plays a crucial role as a tool to measure students' understanding of the material being taught. However,

in practice, some test items still do not conform to the principles of developing good evaluation instruments, such as items that are too easy, too difficult, or unable to differentiate between students with different levels of mastery of the material. This situation indicates the need for further analysis of test item quality, particularly in terms of difficulty level and discrimination power, so that evaluation instruments can truly represent learning outcomes objectively and fairly.

What factors influence those outcomes. For example, research shows that students' socioeconomic backgrounds can influence the effectiveness of learning interventions, with students from disadvantaged backgrounds tending to face greater difficulties in achieving expected outcomes (Ernawati, 2022). Test item quality can be analyzed using a psychometric approach through difficulty level and discrimination power. These aspects indicate how well items measure student ability and distinguish performance levels, ensuring the validity and reliability of evaluations on key topics such as the Japanese Occupation and Indonesian Independence. Therefore, this study examines whether the evaluation instruments meet appropriate quality standards.

Several recent studies have examined the importance of analyzing test item quality in history teaching at the secondary education level. Furthermore, Wati (2022) and Verawati (2023) analyzed test items from various subjects at the secondary school level, focusing on the aspects of difficulty and discriminating power. Their findings showed that unbalanced difficulty levels reduce the effectiveness of evaluation, while low discriminating power makes questions unable to distinguish students' mastery levels, resulting in less valid assessment outcomes. These results highlight the importance of item analysis, including in history, to ensure that evaluation instruments measure students' abilities and depth of understanding accurately.

One crucial element that is often the focus of analysis is the extent to which the test items align with the applicable curriculum. Utomo (2019) emphasized the importance of assessing the validity of question content in the implementation of Mid Semester Assessments and Final Semester Assessments. When the analysis process is carried out unsystematically, the evaluation results have the potential to not accurately represent student learning achievement. Meanwhile, research by Marlina et al. (2024) revealed that the process of identifying and analyzing questions before use can produce evaluation instruments with high validity and reliability, and can serve as a reference for developing better questions in the future.

There is a visible gap between recent research on item analysis, particularly in terms of discriminability and difficulty, and the empirical and theoretical understanding of these studies. Many have explored analytical techniques, such as the use of Item Response Theory and statistical methods, but comprehensive reports on educators' integration of these analysis results into actual classroom practice are lacking. Research by Marlina et al. (2024) and Muhamad Yunus et al. (2021) demonstrates the importance of the item identification process for ensuring assessment quality. However, while analysis can provide feedback for improvement, the implementation of the obtained suggestions remains inconsistent with classroom

practice, where many educators are untrained or lack the capacity to implement them.

Furthermore, while the literature indicates that item analysis for validity and reliability is a crucial step in educational evaluation, there is still a gap in demonstrating how the results of such analysis can be applied to the development of better evaluation tools and in everyday learning. Research by Supandi & Farikhah (2016) emphasizes the importance of distractor function values and the relationship between discrimination and difficulty levels, but lacks practical guidance for teachers in their specific contexts. Many studies, including Yusuf's (2024) study describe analysis procedures without in depth explanations of how teachers can translate these findings into concrete actions in the classroom.

This study aims to examine the difficulty level and discriminatory power of history items on the topic of the Japanese Occupation and Indonesian Independence at the high school level. The main objective of this study is to empirically evaluate the quality of history test instruments to determine the extent to which the test items are able to measure student learning outcomes fairly and proportionally. The novelty of this research lies in its focus on national history material which has high contextual value but is still rarely analyzed in depth in question evaluation studies. In addition, this study also links the analysis findings with their implications for the practice of preparing questions by teachers, so that it is expected to be able to contribute to the development of higher quality and more targeted evaluation instruments in the context of history learning in schools.

2. Methodology

Instrument

This study used a quantitative approach with a descriptive design. The quantitative approach was chosen because the primary objective of the study was to analyze numerical data in the form of difficulty levels and item discrimination, calculated based on student test results. This quantitative approach was used to describe the quality of the test items objectively and systematically without manipulating variables. This study employed a quantitative approach with a descriptive design. This approach was chosen because the primary focus of the study was the collection and analysis of numerical data, specifically the difficulty levels and discriminatory power of test items obtained from student test results. In this approach, the researcher did not manipulate the learning variables but rather described the existing phenomena systematically and objectively (Wulandari et al., 2023).

This research was conducted at SMA Negeri 7 Malang in May 2025. The selection of these schools was carried out purposively by considering a number of strategic factors. One consideration was the accessibility of the school location, which facilitated coordination and direct implementation of research activities. Furthermore, SMA Negeri 7 Malang is known for its well organized learning

administration system, including documentation of evaluation questions and student learning outcomes that can support the data analysis process. The school's readiness to support academic activities, particularly through the openness of history teachers to research collaboration, was also a key factor in selecting this location. With support from the school and teachers, it is hoped that the data collection process, test administration, and validation of findings will run smoothly and according to procedures.

The research subjects were 26 students in grade XI 4. The selection of this class was done purposively by considering that class XI 4 had received material on the Japanese Occupation and the Proclamation of Indonesian Independence according to the academic calendar. Furthermore, the history teacher in this class used evaluation instruments relevant to the material, allowing for an analysis of the quality of the questions. The number of students in this class was also deemed sufficient to conduct a simple statistical analysis of the difficulty level and the discriminating power of the items. The instrument used in this study was a test that had previously been implemented by the history teacher as a learning evaluation tool. The use of this instrument enabled the researcher to analyze the quality of the items, including the level of difficulty and discriminating power . Furthermore, the number of students in this class was deemed sufficient to support a simple statistical analysis of the characteristics of the questions used.

Data collection

Data collection in this study was conducted through an evaluation test using Google Forms. The evaluation instrument consisted of 20 multiple choice questions and 5 essay questions, designed to measure students' understanding of the historical material regarding the Japanese Occupation and the Proclamation of Indonesian Independence. The use of Google Forms was chosen due to its efficiency in reaching all students simultaneously, ease of documenting results, and flexibility in administering the test without the need for printed media. The results of student responses were automatically recorded in the form of quantitative data, which was then used as the basis for analyzing the level of difficulty and the discrimination power of the test items.

Data analysis

Data obtained from the multiple choice questions were analyzed using the latest version of SPSS statistical software. The analysis focused on two main aspects: the difficulty level and the discriminatory power of each item. The difficulty level was calculated to determine the extent to which the item was easy or difficult for students, while the discriminatory power was used to measure the item's ability to distinguish between students with high and low understanding of the material. The calculation was carried out by dividing the student group into two parts, namely the upper group and the lower group based on the total score. Next, the proportion of students from each group who answered each question correctly was compared to determine the question quality category. The results of this analysis are then

interpreted to provide input on the quality of the evaluation instrument and as a basis for compiling more effective questions in the future.

3. Results and Discussion

Results of Question Difficulty Level

The level of difficulty of questions is an important element in educational item analysis, which shows how difficult or easy a question is for students. In an educational context, a deeper understanding of this difficulty level can help teachers design questions that better suit students' abilities and knowledge, and improve teaching effectiveness. Understanding the difficulty level of questions not only aids assessment but also plays a crucial role in curriculum development and teaching methods (Lestari et al., 2023). The results of this study involved 26 grade XI 4 students at SMA Negeri 7 Malang. The primary focus of this analysis is to measure the difficulty level of each test item to determine the extent to which the test items are able to proportionally assess student abilities. Data processing was performed using SPSS version 27 software, using statistical indicators to group the test difficulty levels into three categories: difficult, medium, and easy. This analysis aims to provide a quantitative overview of the quality of the test items and serve as an evaluative basis for developing better assessment instruments in history learning. The results of the test on the level of difficulty of the questions can be seen in table 1. The following are the results of the test on the level of difficulty of the questions.

Table 1. Results of the Multiple Choice Question Difficulty Level Test

Question No.	Question	Difficulty Level	Criteria
1.	The main reason Japan undertook the Meiji Restoration	1.00	Easy
2.	The relationship between Japan's modernization since the Meiji Restoration and Japan's involvement in the Axis Powers	0.96	Easy
3.	Important Japanese figures who played a role in the Axis Powers	0.85	Easy
4.	Strategic reasons for the Japanese occupation of Tarakan	0.27	Difficult
5.	Contents and meaning of the Kalijati Agreement	0.77	Easy
6.	Division of Japanese administrative areas in Indonesia (Rikugun and Kaigun)	0.69	Currently
7.	Romusha policy	0.88	Easy
8.	The aims of Japanese military policy in Indonesia	0.92	Easy

9.	Putera (people's power center)	0.77	Easy
10.	Tonarigumi	0.88	Easy
11.	PETA's role in the history of Indonesian independence	0.96	Easy
12.	MIAI and Masyumi	0.73	Easy
13.	Singaparna resistance	0.81	Easy
14.	PETA Resistance in Blitar 1945	0.81	Easy
15.	The impact of Japanese policy on Indonesian nationalism	0.81	Easy
16.	The impact of Japanese military cadre formation	0.92	Easy
17.	The strategic meaning of BPUPKI	0.88	Easy
18.	Results of the first session of BPUPKI	0.85	Easy
19.	The Role of the Committee of Nine	0.96	Easy
20.	The role of the PPKI	1.00	Easy

Each question item was analyzed using the mean value of the students' work, which was then classified into three categories of difficulty level: difficult (mean 0.00–0.30), moderate (mean 0.31–0.70), and easy (mean 0.71–1.00). This table contains information regarding the question number, the mean value of each question, and the difficulty level category. This data presentation aims to provide a systematic overview of the extent to which the questions are able to test students' competencies proportionally, as well as to identify which questions need to be improved to improve the quality of the learning evaluation instrument. The criteria for the level of difficulty of the questions used in this study can be seen in Table 2. The following table shows the criteria for the level of difficulty of the questions.

Table 2. Criteria for Question Difficulty Level

Interval	**Criteria**
0.00 - 0.30	Difficult
0.31 - 0.70	Currently
0.71 - 1.00	Easy

Based on the results of statistical analysis using SPSS version 27 on 20 multiple choice questions (questions 1 to 20), the difficulty level of each question was calculated by referring to the classification criteria of the average value (mean): values between 0.00–0.30 are categorized as difficult questions, 0.31–0.70 as medium questions, and 0.71–1.00 as easy questions. From the results of the analysis, it is known that 18 questions are included in the easy category, namely question 1, question 2, question 3, question 5, question 7, question 8, question 9, question 10, question 11, question 12, question 13, question 14, question 15, question 16, question 17, question 18, question 19, and question 20. Meanwhile, only one question is classified as medium, namely question 6 with a mean value of 0.69. One question, question 4, was considered difficult, with a mean score of 0.27. This finding indicates that most of the questions were considered easy for students,

with a percentage reaching 90% of the total questions. On the other hand, questions with medium and difficult levels of difficulty only cover 5% each. These findings can be seen in Table 3 and are an early indicator that improvements need to be made in the preparation of test items so that they can accommodate variations in students' ability levels more proportionally. The following are the results of the difficulty level of the descriptive questions.

Table 3. Results of the Level of Difficulty of Essay Questions

Question No.	Question	Difficulty Level	Criteria
1.	The reasons for changing the Jakarta Charter before it was ratified as Pancasila	0.87	Easy
2.	The role of Pancasila in shaping the identity of the Indonesian nation	0.64	Currently
3.	The role of young people and old people in the Proclamation of Indonesian Independence	0.90	Easy
4.	The meaning of the Rengasdengklok incident	0.64	Currently
5.	Admiral Maeda's role in the process of Indonesian independence	0.70	Currently

Based on the classification of the difficulty level of the essay questions, the results show that questions 1 and 3 are included in the easy category, with mean values of 0.87 and 0.90, respectively. This indicates that most students are able to answer these questions correctly. Meanwhile, questions 2 and 4 are classified as medium, with the same mean value of 0.64. This means that these questions have a moderate level of difficulty and are able to test students' understanding proportionally. Question 5 is at the upper threshold of the medium category with a mean value of 0.70, which indicates that although it is classified as medium, its level of difficulty tends to approach the easy category. This analysis shows that in general, these questions have a fairly even distribution of difficulty levels, although special attention is needed to maintain a balance between easy, medium, and difficult questions for more optimal evaluation.

Results of the Differential Power Test of Questions

Discriminant power is a measure or indicator that shows how well a test item can differentiate between test takers with high ability and those with low ability (Azzahroh et al., 2022). Simply put, discriminant power indicates the ability of a question to differentiate students who have truly mastered the material from students who have not mastered it. Questions with high discriminant power will be answered correctly more often by high-achieving students, and answered incorrectly more often by low-achieving students. Discriminant power is defined as the ability of a question to differentiate students based on their cognitive abilities (Muchtar et al., 2024). As a reference in determining the category of the level of question discrimination power, the criteria used are shown in table 4. The following table of criteria for question discrimination power.

Table 4. Criteria for Differentiating Power of Questions

Distinguishing Power	Category
Negative	Not good
0.00 - 0.20	Bad
0.21 - 0.40	Enough
0.41- 0.70	Good
0.71- 1.00	Very good

Item validity was analyzed using the Corrected Item-Total Correlation value, which measures the extent to which an item correlates with the total score of the entire test. This value indicates how well the item is able to measure the same competency as measured by all evaluation instruments. Interpretation of this correlation value is divided into several categories, namely negative values indicate invalid or poor items. Values between 0.00 and 0.20 indicate low or poor validity, while values 0.21 to 0.40 indicate sufficient validity. Values between 0.41 and 0.70 describe good validity, and values 0.71 to 1.00 indicate excellent validity (Hendryadi, 2021). Thus, this validity analysis is important to ensure that each item is able to provide an accurate picture of the material mastery by students. The results of the discrimination power of multiple choice questions can be seen in Table 5. The following are the results of the discrimination power of multiple-choice questions.

Table 5. Results of Differential Power of Multiple Choice Questions

Question No.	*Corrected Item-Total Correlation* Value	Validity Category
Question 1	0.282	Enough
Question 2	-0.015	Not good
Question 3	0.547	Good
Question 4	-0.256	Not good
Question 5	0.375	Enough
Question 6	0.299	Enough
Question 7	0.439	Good
Question 8	0.529	Good
Question 9	0.583	Good
Question 10	0.310	Enough
Question 11	0.599	Good
Question 12	0.597	Good
Question 13	0.324	Enough
Question 14	0.085	Bad
Question 15	0.542	Good
Question 16	0.477	Good
Question 17	0.708	Very good
Question 18	0.389	Enough
Question 19	0.599	Good
Question 20	0.000	Bad

There are two questions, namely question 2 and question 4, which are declared invalid because they have a negative correlation value. In addition, two other questions, namely question 14 and question 20, showed relatively poor validity. A total of six questions have validity in the sufficient category, while nine questions show good validity. Only one question, namely question 17, falls into the very good validity category. Overall, most of the test items demonstrated adequate validity

and were reliable in measuring student abilities. However, further evaluation of test items 2 and 4 is warranted. These items should be revised or even removed, as they did not contribute positively to the total score of the evaluation instrument. The results of the discriminant power of the essay questions are presented in Table 6. The results of the discriminant power of the essay questions are as follows.

Table 6. Results of Differential Power of Essay Questions

Question No.	*Corrected Item-Total Correlation* Value	Validity Category
Question 1	-0.078	Not good
Question 2	-0.100	Not good
Question 3	-0.277	Not good
Question 4	0.382	Enough
Question 5	0.205	Bad

Based on the analysis results using SPSS 27, it can be concluded that questions 1 to 3 are included in the poor category because they have a negative Corrected Item-Total Correlation value, which means these items are not positively correlated with the total score and are not valid as a measuring tool. A negative item-total correlation indicates that the item is inversely correlated with the total score, meaning that participants with high total scores tend to answer the item incorrectly. This generally indicates problematic items (wrong mis-keys, ambiguity, or too strong trap options) (Zijlmans et al., 2018). Furthermore, question number 4 shows validity in the sufficient category, which means it still has a positive correlation although not too strong. Meanwhile, question number 5 is included in the poor validity category because its correlation value is below 0.20 . These findings indicate that several items need to be revised to improve the overall quality of the evaluation instrument.

The results of the difficulty analysis of 20 multiple-choice questions given to 26 grade XI 4 students at SMA Negeri 7 Malang showed that most of the multiple-choice questions were classified as easy, with a percentage reaching 90% (18 questions). Only one question was in the medium category, and another was in the difficult category. These findings indicate that most students were able to answer these questions correctly, which reflects that these questions were less challenging and less able to test students' understanding in depth. In the context of evaluating history learning, especially on the Japanese Occupation and Indonesian Independence material, these results indicate that the question instruments are still not balanced in terms of the level of difficulty. Questions that are too easy have the potential to produce inaccurate data in differentiating students' levels of mastery. In other words, the instrument's validity as a measure of ability is suboptimal. This situation does not fully support the goals of formative and summative evaluation, which ideally should depict an even distribution of student abilities.

In addition, the results of the analysis of descriptive questions also show a dominance of questions in the easy category (2 questions), while the rest are in the medium category. The absence of questions that are classified as difficult shows a similar tendency to multiple choice questions. This means that both the multiple

choice and descriptive forms do not reflect sufficient variation in levels of difficulty. In a broader perspective, these findings emphasize the importance of designing evaluation instruments that take into account the proportion of questions of varying difficulty levels. This is important not only for measuring overall cognitive achievement but also for increasing the challenge of questions and encouraging students to think critically. Overall, these findings reinforce the urgency of developing more balanced questions so that assessments can function optimally in the learning process. Teachers should regularly evaluate the quality of test items as part of their learning reflection and assessment improvement efforts.

Discriminant power analysis of 20 multiple choice questions show variations in the ability of the questions to differentiate between high and low ability students on the Japanese Occupation and Indonesian Independence topic. Discriminant power is an important indicator of the quality of a question, as it reflects the extent to which the item is able to identify differences in the level of mastery of the material between students. In this finding, two questions, namely questions 2 and 4, have a negative correlation value, which means they are not able to differentiate students well. The negative correlation in the context of discriminant power indicates that students who received high scores overall actually answered the questions incorrectly, while students with low scores answered correctly. This indicates an error in the question construction, such as inaccurate answer options, ambiguity of statements, or inconsistency with competency indicators. Because they do not provide a positive contribution to mapping student abilities, questions 2 and 4 should be revised completely or removed from the evaluation instrument.

Furthermore, two other questions, question 14 and question 20, had relatively poor discriminatory power. While not as bad as the negative correlation, these questions still demonstrate weaknesses in identifying differences in ability between students. Six questions had sufficient discriminatory power, indicating that they had moderate discriminatory power but still needed improvement. Nine questions showed good discriminatory power, and one question (question 17) was in the excellent category. This indicates that most of the questions were able to carry out their discriminatory function quite effectively.

Overall, these results indicate that the quality of the discriminating power in the evaluation instruments is relatively varied, with a positive trend for most items. However, the presence of items with low and negative discriminating power is a serious concern (Cappelleri et al., 2015). Therefore, continuous evaluation of the discriminating power of items is necessary, especially in the context of formative and summative assessments, so that the evaluation truly reflects student learning achievement. Teachers are expected to make improvements to items with low discriminating power or invalid, in order to improve the quality of objective, accurate, and fair evaluations.

Discriminatory power analysis of five descriptive questions indicate that the instrument's discriminatory quality still needs improvement. Questions 1 to 3 were found to have negative Corrected Item-Total Correlation Values. A negative

correlation indicates that the correct answers to the item are not positively correlated with the student's total score, so these questions do not function as intended in distinguishing between high and low ability students. Thus, these items are categorized as inadequate and invalid as measuring tools in learning evaluation. The presence of such questions can degrade the overall quality of the instrument and distort the assessment of learning outcomes (Quaigrain & Arhin, 2017).

Question number 4 shows a moderate level of discrimination. Although the correlation isn't high, this positive value indicates that the question still has moderate discriminatory power. Questions like this can still be used in evaluations, but it is recommended that improvements be made to the wording and accuracy of the indicators to improve their quality. Meanwhile, question number 5 is categorized as having poor discriminatory power, as its correlation value is below the minimum threshold (0.20). This indicates that although it still has a positive correlation, the question's ability to differentiate students with high and low mastery of the material is relatively low. This item needs to be reviewed, both in terms of content, sentence structure, and suitability to learning objectives. Overall, these results emphasize the importance of discriminatory power analysis in developing quality evaluation instruments. A good instrument must not only measure students' general abilities but also be able to accurately differentiate students' levels of mastery. Revision of questions with low or negative discriminatory power is an important step in ensuring the validity and reliability of learning evaluations.

In the context of research on the discriminating power and difficulty of questions at the secondary school level, several relevant studies can provide additional insights. Research conducted by Saleha et al. (2022) provides a quantitative analysis of the final exam for Indonesian language at SMP Negeri 3 Kutacane. Their findings indicate that 55% of the questions tested fall into the intermediate category, while 42.5% fall into the difficult category, and only 2.5% fall into the easy category. This provides an overview of the distribution of question difficulty in the secondary school context that can be applied to other subjects such as history. Similarly, research by Zulfiana et al. (2023) provides an analysis of the validity, reliability, discriminating power, and difficulty of questions in science learning. Although the primary focus of this study is on science, the methodology can be adapted to history. The results show that the instruments used have high validity and reliability, which are important indicators to ensure the quality of questions in education.

Another study conducted by Faridah (2021) highlighted the characteristics of test items in the final semester history assessment for 11th-grade students. The results showed that the majority of the analyzed questions met the eligibility standards. This study also emphasized the importance of evaluating the discriminating power and difficulty level of history questions to ensure that they accurately measure students' understanding and knowledge of the historical context being taught. The implications of these findings suggest that in developing evaluation instruments, particularly in history, teachers need to focus not only on developing materially appropriate questions but also on the psychometric quality of the questions. In other words, statistical analysis of discriminatory power and difficulty levels is crucial to

ensuring that the questions are able to differentiate students' levels of understanding fairly and proportionally. This contributes to more meaningful learning and more objective evaluation, allowing the results to be used as a basis for making appropriate educational decisions, both in improving teaching methods and planning remedial measures.

This study has several limitations. One of the main limitations lies in the quality of the items analyzed. Based on the results of the evaluation of the level of difficulty and discrimination, it was found that most of the items were too easy and unable to differentiate students' abilities effectively. This indicates that many items are still inadequate and need to be revised or even completely replaced to function as a valid and reliable evaluation tool. Furthermore, limitations also lie in the number and characteristics of the sample used, which only involved 26 students from one class in one school, so the results of this study cannot necessarily be generalized to a wider population. Another limitation is the focus of the analysis, which only covers the level of difficulty and discrimination, without considering other aspects such as content validity, item construction, or the linkage of the items to the overall learning indicators. Given these limitations, it is hoped that further research can expand the scope of the analysis, use larger and more diverse samples, and refine the item instrument so that learning evaluation becomes more meaningful and accurate.

4. Conclusion

This study revealed that most of the test items on the Japanese Occupation and Indonesian Independence subject were relatively easy and had varying levels of discrimination, including several items with negative discrimination values. These findings indicate that the assessment instrument used still requires significant improvement to be able to measure students' abilities validly and reliably. Thus, the results of this study emphasize the importance of quantitative item evaluation as an integral part of the process of developing and revising history assessment instruments.

However, limitations of the study, such as the sample size being limited to one class and the analysis focusing only on difficulty level and discrimination power, need to be considered. Therefore, these findings cannot be broadly generalized without further research with a larger scope and more comprehensive methods. Overall, this research contributes to strengthening understanding of the importance of empirical analysis in evaluating question quality, particularly in the context of history learning. By refining the instrument based on these results, it is hoped that assessments will be more effective in reflecting student competencies and supporting a more meaningful and focused learning process.

References

Azzahroh, S., Iman, F. L., Anwar, B., Islam, U., & Malik, M. (2022). Analisis butir

soal ujian akhir semester mata kuliah psikologi belajar menggunakan software Anates. *Journal of Indonesian Psychological Science*, *03*(2).

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2015). Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing PatientReported Outcome Measures. *Science*, *23*(1), 1–7. https://doi.org/10.1016/j.clinthera.2014.04.006.Overview

Ernawati, F. (2022). Penjaminan Mutu Calon Guru Anak Usia Dini di Perguruan Tinggi Kota Surakarta. *Jurnal Obsesi : Jurnal Pendidikan Anak Usia Dini*, *6*(6), 6296–6308. https://doi.org/10.31004/obsesi.v6i6.3245

Faridah, A. (2021). Karakteristik Butir Soal Penilaian Akhir Semester Mata Pelajaran Sejarah Kelas XI SMA Negeri 1 Pangkalpinang. *Fajar Historia: Jurnal Ilmu Sejarah dan Pendidikan*, *5*(2), 210–221. https://doi.org/10.29408/fhs.v5i2.4609

Hendryadi. (2021). Editorial Note: Uji Validitas Dengan Korelasi Item-Total ? *Jurnal Manajemen Strategi dan Aplikasi Bisnis*, *4*(1), 315–320. https://doi.org/10.36407/jmsab.v4i2.404

Lestari, M., Halini, & Indriani, T. (2023). Analisis Tingkat Kesukaran Soal Persamaan dan Pertidaksamaan Nilai Mutlak Melalui Pendekatan Teori Tes Klasik. *1*(2). https://journal.uns.ac.id/ijolii

Marlina, M., Rahim, A., Aziz, F., & Arsalaan, A. T. (2024). Pelatihan analisis kualitas instrumen penilaian hasil belajar dengan pendekatan Clasical Theory dan Item Response Theory di sekolah dasar. *KACANEGARA Jurnal Pengabdian Pada Masyarakat*, *7*(2), 209. https://doi.org/10.28989/kacanegara.v7i2.1888

Muchtar, Z., Gusfa, M., Dibyanti, R. E., Sutiani, A., & Sinaga, M. (2024). Pengembangan Instrumen Evaluasi untuk Mengukur Keterampilan Berpikir Tingkat Tinggi pada Materi Laju Reaksi. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *7*(8), 8149–8155. https://doi.org/10.54371/jiip.v7i8.4901

Muhamad Yunus, Lalu Wirajayadi, Neni Suryanirmala, Aliahardi Winata, & Zul Haeri. (2021). Pkm Peningkatan Kualitas Hasil Evaluasi Pembelajaran Siswa Menggunakan Analisis Butir Soal dengan Program Iteman dan Spss Di Desa Jago Kecamatan Praya Kabupaten Lombok Tengah Provinsi Nusa Tenggara Barat. *J-ABDI: Jurnal Pengabdian Kepada Masyarakat*, *1*(3), 399–408. https://doi.org/10.53625/jabdi.v1i3.157

Quaigrain, K., & Arhin, A. K. (2017). Using Reliability And Item Analysis to Evaluate a Teacher-Developed Test in Educational Measurement and Evaluation. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1301013

Saleha, James Marudut, & Rekaza Akbar. (2022). Analisis Butir Soal Ujian Akhir Sekolah Mata Pelajaran Bahasa Indonesia Kelas VIII SMP Negeri 3 Kutacane Tahun Pelajaran 2020/2021. *Tuwah Pande: Jurnal Ilmu Pendidikan Dan Pengajaran*, *1*(2), 254–269. https://doi.org/10.55606/tuwahpande.v1i2.25

Supandi, S., & Farikhah, L. (2016). Analisis Butir Soal Matematika Pada Instrumen Uji Coba Materi Segitiga. *JIPMat*, *1*(1), 71–78. https://doi.org/10.26877/jipmat.v1i1.1085

Utomo, B. (2019). Analisis Validitas Isi Butir Soal sebagai Salah Satu Upaya

Peningkatan Kualitas Pembelajaran di Madrasah Berbasis Nilai-Nilai Islam. *Jurnal Pendidikan Matematika (Kudus)*, *1*(2). https://doi.org/10.21043/jpm.v1i2.4883

Verawati, Y. (2023). Analisis Butir Soal Ujian Akhir Semester (UAS) Mata Kelas VII SMP Islam At Tanwir Kecamatan Ledokombo. *3*(01), 114–121. https://doi.org/10.57008/jjp.v3i01.422

Wati, E. (2022). Analisis Butir Soal Pilihan Ganda Mata Pelajaran Bahasa Indonesia Kelas VII Mts Attaufiqiyah. 1–7. https://doi.org/10.21070/ups.9107

Wulandari, E., Faturrohman, H., Widodo, S. T., Wahyuni, N. I., & Ningsih, F. (2023). *Pengaruh Penggunaan Media Interaktif Terhadap Motivasi Belajar Peserta Didik Mata Pelajaran Pendidikan Pancasila Kelas Ii Sdit Insan Mulia Semarang*. *22*(2), 19–25. https://journal.stkipsubang.ac.id/index.php/didaktik/article/view/2086/1739

Yusuf, F. W. (2024). Analisis Butir Soal Asesmen Sumatif Biologi Materi Perubahan Lingkungan Dengan Menggunakan Anates Pada Kelas X Sma. *LEARNING : Jurnal Inovasi Penelitian Pendidikan Dan Pembelajaran*, *4*(1), 126–135. https://doi.org/10.51878/learning.v4i1.2777

Yuwono, I., & Mirnawati, M. (2021). Strategi Pembelajaran Kreatif dalam Pendidikan Inklusi di Jenjang Sekolah Dasar. *Jurnal Basicedu*, *5*(4), 2015–2020. https://doi.org/10.31004/basicedu.v5i4.1108

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2018). Item-Score Reliability in Empirical-Data Sets and Its Relationship With Other Item Indices. *Educational and Psychological Measurement*, *78*(6), 998–1020. https://doi.org/10.1177/0013164417728358

Zulfiana, S., Gunamantha, I. M., & Putrayasa, I. B. (2023). Pengembangan Instrumen Kemampuan Berpikir Tingkat Tinggi Dan Literasi Sains Pada Pembelajaran Ipa Kelas V Sd. *PENDASI: Jurnal Pendidikan Dasar Indonesia*, *7*(1), 13–24.

How to cite this article:

Putri, A. P., & Sayono, J. (2026). Evaluation of Item Quality: Analysis of Difficulty Level and Distinction Power with Quantitative Methods. *Journal of Educational Sciences*, *10(1)*, 317-330.