IJOLAE Indonesian Journal on Learning and Advanced Education

http://journals.ums.ac.id/index.php/ijolae

ITEMAN-Based Evaluation of End-of-Semester Assessment Items: A Case Study of Language Test in Indonesian School Context

Riswanda Himawan[1]✉, Hermanto[2], Burhan Nurgiyantoro[3], Suyono[4], Didin Widyartono[5], Agustina Purwanti[6], Le Yujing[7], Victor A. Pogadaev[8]

[1,4,5]Faculty of Letters, Universitas Negeri Malang, Indonesia
[2]Faculty of Teacher Training and Education, Universitas Ahmad Dahlan, Indonesia
[3]Faculty of Languages, Arts and Culture, Universitas Negeri Yogyakarta, Indonesia
[6]Sekolah Menengah Pertama Negeri 8, Kota Yogyakarta, Indonesia
[7]Faculty of Education, University Hainan College of Foreign Studies, China
[8]Faculty of Languages and Linguistics, Moscow State Institute of International Relations, Russia

DOI: 10.23917/ijolae.v6i3.23254
Received: December 22nd, 2023. Revised: May 31st, 2024. Accepted: August 27th, 2024
Available Online: September 20th, 2024. Published Regularly: September, 2024

Abstract

Assessment is an inseparable part of the learning process. Assessment is a way to decide the choices that a teacher will make for students, so that what they do is truly measurable. In line with this description, this study aims to determine the quality of the end-of-semester assessment questions for Indonesian, grade VIII, at SMPN 8 Yogyakarta. The method used in this research is descriptive quantitative. Quantitative descriptive is used to analyze documents in the form of output analysis results from the ITEMAN application. The results of the review that will be submitted later are in the form of validity, reliability, IDB, ITK and the functioning of the distractor. In obtaining the results of the study. The research steps carried out in this study were are testing the validity of the question grids, testing the reliability, IDB, ITK, and distractors through the ITEMAN program, analyze the results according to expert opinions and research that are relevant to this study, and conclude the results of the study. There were 50 questions that were analyzed and tested on 32 class VIII students of SMPN 8 Yogyakarta. Overall, the research results show all questions are declared valid, Alpha is 0.926. This shows that the reliability of the items is very high, IDB results show 38 items in good, 5 moderate, 4 sufficient, and 3 failed categories, the results of ITK analysis showed 31 very easy questions, 10 easy questions, 5 moderate questions, 2 difficult questions, and 2 very difficult questions, and the results of the distractor function show that 32 questions do not work, and 18 item distractors work. Overall, the results of this analysis aim to provide a reference base related to the quality of questions through ITEMAN, especially in learning Indonesian, which so far has not been done much.

Keywords: assesment item semester, indonesian school context, ITEMAN analysis, item discrimination index, learning objectives achievement

✉***Corresponding Author:***
Riswanda Himawan, Faculty of Letters, Universitas Negeri Malang, Indonesia
Email riswandahimawan.242119@students.um.ac.id

1. Introduction

Assessment is an inseparable part of the learning process (Asrial et al., 2023; Prastikawati et al., 2024). Assessment needs to be designed in such a way that learning objectives can be achieved (Kusumaningtyas et al., 2024).

It serves as a strategic tool for teachers to determine the most effective learning activities and ensure student progress is truly measurable

(Susanto et al., 2015). One of the assessment types at the middle level is the end-of-semester or final exam. This exam serves as a process for both teachers and students to evaluate learning outcomes achieved throughout the semester. It allows the teachers and students to reflect on the results, so that adjustments can be made to improve future learning experiences. (Nurhalimah et al., 2022)

At the end of the semester, students receive grades compiled with scores from other assessments (Ruay Garcés, 2018). These grades are measured by certain achievement standards, serving as parameters for student progress and learning success (Wahyuni & Kurniawan, 2018). This is in line with Kurniawan (2015) who stated that the final assessment is included in a summative test, which functions to determine the extent to which students have achieved competence in certain subjects. The results from these tests are then compared to learning objectives or the minimum passing grade (KKM) (Muhith, 2018).

Therefore, the results of the end-of-semester assessment serve a dual purpose: they function as a record of student learning progress and determine a student's eligibility for the next program. In this sense, the assessment falls under the category of learning evaluation (Anggraini & Suyata, 2014; Parancika & Suyata, 2020; Alnovgada & Suyata, 2019).

Assessment or evaluation of learning is very closely related to test instruments (Timor, 2022). It is an important instrument used to measure learning achievement. A test can be defined as an assessment tool that uses questions or instructions for student to answer and complete. (Mania et al., 2020).

Purniasari, Masykuri, & Ariani (2021) stated that a valid instrument must be thoroughly tested to effectively evaluate student learning outcomes across all domains, from character development to critical thinking skills. A test is defined as an instrument or systematic procedure used to observe and measure one or more student characteristics. This measurement is typically done using a numerical scale or a classification scheme (Nitko & Brookhart via Iskandar & Rizal, 2018).

This description emphasizes the importance of developing tests following proper test guidelines. Before using the test, the teacher must test the validity of the test instrument to meet the requirements for both validity and reliability. As a part of learning, evaluation is one of the most significant components (Purniasari et al., 2021). However, despite its undeniable importance, the role of evaluation in facilitating learning outcomes is not always fully recognized, especially teachers as facilitators in learning (Himawan & Nurgiyantoro, 2022).

This is further supported by the results of initial observations in several schools, which indicate that the education system included learning, must be balanced with good assessment. Many teachers have difficulty processing student assessment scores. Teachers often assign grades directly to students without analyzing the question items, potentially overlooking the crucial principles of clear learning evaluation (Fitriani et al., 2020); Azizah & Sumardi, 2021)

Building on this point, this align with Rotama, Budiutomo, & Bowo (2020) who identify several issues within Indonesia's education evaluation system. They point out an overemphasis on students' cognitive abilities. Additionalyy, the instruments used are very limited. Develop by teachers, they lack essential validation and reliability testing, as well as crucial item analysis processes like discrimination index, difficulty level, and distractor functioning analysis (Fridaram et al., 2021). As Himawan, Suyata (2024) suggest, a valid assessment instrument for learning evaluation necessitates both high validity and reliability (Arifin & Retnawati, 2017).

Item analysis can be conducted using two primary theories: Classical Test Theory (CTT) and Item Response Theory (IRT). This study employed CTT to analyze the instrument's results through the ITEMAN program. A widely used software program, ITEMAN is designed for classical item analysis (Alfarisa et al., 2019). It is part of the MicroCATn software suite developed by the Assessment Systems Corporation in 1982 (Himawan & Nurgiyantoro, 2022)

Subsequently, Himawan & Nurgiyantoro, (2022) explain that there are several stages to carrying out item analysis, using the ITEMAN computer program. The data input process begins by creating a text file. User can navigate from the Start menu and searching for, "Notepad". This file serves as the input for the ITEMAN program and requires specific information: The first line should include the number of questions in the assessment, a code for omitted responses (typically "O" or "0") for items unanswered questions, the population code (denoted by "N"), and the type of participant identification number (usually a number). The second line requires the answer key which can be filled with answer options (e.g., A, B, C, D, E). The third line specifies the total number of answer options available for the questions (Nanda Pratiwiningtyas et al., 2017). The fourth line allows users to request analysis for each item. "Y" indicates the item should be analyzed, "N" not to be analyzed. Finally, the fifth line requires the student answer alongside their corresponding identification information. Next, it is important to remember to save the data file within the same folder where the ITEMAN program is located (Himawan & Suyata, 2022; Arvianto, 2016; Shanta Monica, 2013)

Nurgiyantoro (2016) explains that the results of the ITEMAN analysis consist of Item Statistics and Alternative Statistics. The former (statistics for items) consists of Seq No. (sequential number) according to the order of data entry. The scale Item is the serial number of the item. Pop. Correct (proportion of correct answers/difficulty level) contains an index of the proportion of correct answers per item which shows the item difficulty index. Biser is the biserial correlation between the correct answer per item and the correct answer score. Point Biser is the point biserial correlation between the correct answers per item and the total score (Wijaya et al., 2019; Al-faruq, 2023; Mustafidah et al., 2021). This correlation coefficient is expressed as the discriminating power index (IDB) (Himawan & Nurgiyantoro, 2022).

Validity refers to the degree to which a test instrument measures what it is intended to measure. A test instrument with high validity is appropriate to carry out measurements or data collection, and the results will be precise and accurate (Hanifah, 2014; Setiawan, Susongko, & Hayati, 2020)

Reliability, on the other hand, refers to the consistency of instrument's measurements. A reliable test will produce consistent results when administered multiple times under similar conditions (Mardiana & Suyata, 2017; Nurhalimah et al., 2022)

Building on the importance of validity, (Nurgiyantoro et al., 2020) suggests that the instrument's validity can be assessed through the review indicators. The review includes aspects of material, construction, and language. The material aspect contains matters relating to (a) the conformity of the items with indicators in the blueprint; (b) The suitability of the content of the material with science; (c) the Answer key; (d) the function of the distractor option in the test items. The construction aspect includes (a) the clarity of the formulation of the main problems; (b) the clarity of answer choices; (c) homogeneous answer choices; (d) certainty that there is no double negative form; (e) determining the length of the answer for

each item; (f) there is no dependence between items; (g) the order of the choices in the form of numbers and time (Himawan & Nurgiyantoro, 2022).

Meanwhile, the aspect of language includes (a) communicative language; (b) grammatical sentences; (c) no double meaning of sentences; (d) vocabulary selection/diction. In the discussion of item analysis, classically, through the ITEMAN program, the test items are regarded as feasible if the item difficulty index (ITK) falls within an acceptable range and the discriminating power index (IDB) meets the requirements (Susanto et al., 2015). The item difficulty level index (ITK) shows how easy or how difficult an item is for the test takers, while the discriminating power index (IDB) is a statement about how far an item can differentiate the ability of participants in the high and low groups (Nurgiyantoro, 2016).

An item with a difficulty index (ITK) between 0.20-0.80 is considered acceptable. This range can be further categorized: difficult (0.20-0.40), moderate (0.41-0.60), and easy (0.61-0.80). The discriminating power index (IDB) can be claimed eligible if the index is greater than or equal to 0.20 (Nurgiyantoro, 2016). The function of the right distractor is a good item chosen evenly by students. Conversely, if the items are not effective, students will tend to choose the answer choches unevenly (Putri & Ofianto, 2019)

Furthermore, Nurgiyantoro (2016) identified that there are several criteria for determining the effectiveness of the distractor, namely (1) all distractors (false-options) must be selected, (2) the number of false-option voters from the high group participants must be less than the low group, and (3) if there is only one false-option voter, he must be from the low group. Criteria (2) and (3) are often seen as burdensome, which in essence are similar to the logic of the IDB's demands above, so only Criterion (1) is used effectively. False options were ineffective because none of the test takers chose, as a consequence, the item had to be revised (Magdalena et al., 2021).

Based on the elaboration, the research describe the results of the item analysis for the final exam in the Indonesian language subject at the State Junior High School 8 Yogyakarta. The analysis aimed to find out the result of item analysis for the end semester assessment for Class VIII in Indonesian Language subject, which has been tested in Class VIII at State Junior High School 8 Yogyakarta. In addition, it tried to provide a basis in the form of references to teachers who will carry out the analysis of the test items developed.

In the context of the item analysis using the ITEMAN program, there are several previous studies relevant to the research. The first research was conducted by the first, (Alfarisa et al., 2019) with their research entitled Item Analysis of Social Science Test Using ITEMAN Software for Class V Elementary School. This study has similarities in that both analyze the items using classical theories and the ITEMAN program. In contrast to (Pangesti et al., 2020) who examined social science test items for fifth graders, this study investigated items for the end-of-semester assessment in Indonesian language specifically for Class VIII at State Junior High School 8 Yogyakarta.

The research contribution made by Alfarisa et al (2019) to this study is to provide various theoretical foundations regarding item analysis and a starting point in the form of a method for analyzing test items using the ITEMAN program.

Second, relevant and previous research was conducted by study Setiawan et al, (2022) with the research entitled ”Item Analysis of End-of-Semester Test (PAT) for Indonesian Language Subject in Class XI State Senior High School 1 Polanharjo, Klaten”. This study shares similarities with the current study as it analyzes

test items in the Indonesian language subject using the ITEMAN program.

The difference is that the study (Setiawan et al., 2022) analyzed the items for the senior high school student level. However, this study analyzed the Indonesian language test items for the junior high school level. Concerning the contributions made, the research has produced various findings on item analysis, including validity, reliability, IDB, ITK, and distractor analysis. This study can provide theoretical contributions related to theories on item analysis using the ITEMAN program.

Another relevant research by Purniasari et al. (2021) entitled ”Item Analysis for the Chemistry Subject School Exam at State Senior High School 1 Kutowinangun for the 2019/2020 Academic Year Using the ITEMAN and Rasch Models”.

In common, all research focused on analyzing test items in one subject. Research (Purniasari et al., 2021) analyzes the test items and school exam at the high school level. Meanwhile, this study analyzes the end-of-semester test for the Indonesian language subject at the junior high school level. Research (Purniasari et al., 2021) produced several findings regarding the validity and reliability of the test so that it can contribute to providing a theoretical basis for this study.

Several research related to item analysis using the ITEMAN program have been conducted. This study is a continuation of those research (Anggraini & Suyata, 2014). The aspects that have not been addressed in those research will be the focus of this study, thereby complementing the related research on item analysis with the ITEMAN program and contributing to the evaluation of learning in schools.

Based on the description, this research aims to find out how the final assessment items for the Indonesian language subject at SMPN 8 Yogyakarta are analyzed. It encompasses the validity of the grid as a guideline in creating appropriate and useful grids, determining test reliability which reflects the consistency of an assessment instrument, and understanding the IDB, ITK, and distractors in the evaluation questions developed by the teachers at SMPN 8 Yogyakarta. These components can be used as references for the teachers in developing questions before administering them to students, allowing for an assessment of question quality that can promote critical thinking processes and achieve learning objectives.

Based on these research objectives, the research questions formulated for this research are as follows; (1) what are the results of the validity of the question created by the teacher and applied to students in PAS at SMPN 8 Yogyakarta; (2) What is the reliability of the questions created by the teacher and applied to students in PAS at SMPN 8 Yogyakarta; (3) What are the results of the IDB, ITK and distractor question assessments created by the teacher and applied to students in PAS at SMPN 8 Yogyakarta.

Overall, the research aims to provide a reference base for improving the quality of education, especially at the junior high school level. The novelty demonstrated in this research includes several steps regarding item analysis using the ITEMAN program, which currently needs to be performed. The aspects described in this research will serve as a contribution and reference for conducting item analysis before the questions are administered to students. This approach aims to produce high-quality questions that effectively promote the achievement of learning objectives.

This research, through its focus on item analysis in learning evaluation, aligns with the goals of Putri Pangestu & Rohinah (2019) to enhance the quality of teacher-developed assessment questions. Ultimately, this

improvement contributes to achieving the broader objective of raising educational quality.

2. Method

The research on the item analysis of the final exam for the Indonesian language subject in Class VIII State Junior High School 8 Yogyakarta is included in the document analysis research using a quantitative descriptive approach. Quantitative descriptive analysis was used to describe the results of the score obtained in the ITEMAN program. Quantitative data analysis was chosen as the source of data analysis because it was used to describe data in the form of numbers resulting from the translation of question item analysis, which was viewed through the ITEMAN program.

The quantitative approach in this research was implemented to describe the results of ITEMAN output, which was combined with theory, expert judgement, and several studies relevant to this research. This approach aimed to find data to be discussed regarding item analysis as a reference for teachers in developing evaluation questions in schools, which will undoubtedly have an impact on the advancement of education.

The sources of data information used in this study were the blueprint, test items, and answer sheets for students' answers or responses. Based on this description, the data collection in this study was carried out through documentation, namely documenting the students' answers, related to the test items which were developed. There were 50 questions analyzed and the number of answers analyzed using ITEMAN was taken from 32 students, with various student conditions. There are students with high, moderate, and low abilities. The try-out was conducted at the State Senior High School 8 Yogyakarta in November 2022.

In this study, the analysis of end-of-semester test items for Indonesian Language, Class VIII State Senior High School 8 Yogyakarta was carried out by analyzing the blueprint, questions, and answer sheets containing student responses. The item analysis of the items is seen and reviewed through validity, reliability, level of difficulty, discriminating power, and the effectiveness of distractor with classical theory using the ITEMAN computer program. In addition, this study will also analyze and provide examples of descriptions of test items that are appropriate and not appropriate for use.

The data analysis process included several stages, namely; (1) documenting PAS class VIII questions at SMPN 8 Yogyakarta; (2) measuring the validity of the items by analyzing the blueprint developed according to several indicators from experts including the quality of question writing, writing of words, clauses and sentences to the distribution of cognitive levels contained in the questions; (3) carrying out reliability tests and checking the question difficulty index, discriminating power index, and distractor function using the ITEMAN program, by recording the results of students' answers and then processing them using the ITEMAN program; (4) analyze the results according to expert opinion, relevant theory and research; (5) draw conclusions and classify the quality of the questions and; (6) analyzing items that meet the requirements and do not meet the requirements.

The validity of the test instrument according to (Nurgiyantoro, 2016) was shown by several indicators. The item validity was measured by the researchers' colleagues using the indicators proposed by the experts. The indicators include some aspects, namely material, construction, and language. First, the material aspect consists of (a) the conformity of the items with indicators in the blueprint, (b) the conformity of the items with science, (c) the answer key, and (d) the distractor function. Second, the construction aspect includes: (a) the clarity of item formulation, (b) the clarity of the options, (c) the homogeneous answer choices, (d) no items with double negative statements, (e) the length of

each item, (f) dependency of each item, and (g) the order of items including numbers and time. At last, the language aspect focuses on (a) the communicative language used in the items, (b) grammar, (c) sentences with a double meaning, and (d) the use of vocabulary. These components were used as guidelines for assessing the validity of the grids, ensuring that the questions to be administered to students were appropriate in terms of format, content, number of items, and alignment with indicators (Nurgiyantoro, 2016).

The reliability of the items was measured using ITEMAN by checking the statistics of the Alpha section of the test items. Then, analyzed several items using ITEMAN program. The discrimination index of the items was shown by the point-biserial correlation coefficient. The biserial point was calculated using ITEMAN. The item difficulty level was shown by proportional correct answers calculated using ITEMAN, and the distractor function is shown by proportional endorsing value in the ITEMAN program (Himawan & Nurgiyantoro, 2022)

After obtaining the ITEMAN output, the results were interpreted by expert judgment, including the Point Biserial, which is the point biserial correlation between the correct answers per item and the total score. This correlation coefficient was expressed as the Item Discrimination Index (IDB). The Item Difficulty Index indicates how easy or difficult an item is for the tested participants. The functionality of distractors can be seen through the proportional endorsing value in the ITEMAN program (Nurgiyantoro, 2016). Overall, this research not only involved data presentation and analysis but also drew conclusions from the findings.

3. Result and Discussion

The results of the item analysis, particularly in terms of validity, are presented below:

1. Validity

Validity refers to the degree to which a test measures what it is intended to measure. Data or information can be considered valid if it accurately reflects the actual situation.

Item validity is the appropriateness of test items in relation to the indicators that refer to the definitions or rules of the item before being administered to students (Magdalena et al., 2021); (Retnawati, 2015); Dewi & Sudaryanto, 2020); (Magdalena et al., 2021). Based on these statements, it is evident that item validity must be assessed to determine the fundamental quality of the items. This includes verifying the validity of the item writing, punctuation errors, and the distribution of cognitive levels before the items are presented to the students (Iskandar & Rizal, 2018); (Martin, 2020); (Yadi, 2017).

The validity of the test instrument according to (Nurgiyantoro, 2016) is shown by several indicators. The item validity was measured by the colleagues using the indicators proposed by the experts. The indicators included some aspects, namely material, construction, and language. First, the material aspect consists of (a) the conformity of the items with indicators in the blueprint, (b) the conformity of the items with science, (c) the answer key, and (d) the distractor function. Second, the construction aspect includes: (a) the clarity of item formulation, (b) the clarity of the options, (c) the homogeneous answer choices, (d) no items with double negative statements, (e) the length of each item, (f) dependency of each item, and (g) the order of items including numbers and time.

At last, the language aspect focuses on (a) the communicative language used in the items, (b) grammar, (c) sentences with a double meaning, and (d) the use of vocabulary. These indicators were used as references to assess the validity of the grids against the questions developed by teachers, ensuring alignment

between the items planned in the grids and those written in the evaluation questions.

Based on the explanation above, the item validity analyzed in this research can be further explained as follows.

Table 1. Question Validity Study Results

No	Category	Item	Amount
1	Valid	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50	50
2	In Valid	-	**0**

Validity is the ability of an instrument to measure accurately according to the circumstances to be measured. A test instrument that has a high level of validity is appropriate for assessing and collecting data because the results will be precise and accurate (Syaifudin, 2020). A good test instrument is considered to have high reliability if it can produce data that are relatively consistent (Mardiana & Suyata, 2017). Retnawati (2015) stated that validity and reliability are often discussed in measurement results. Validity is related to the quality of the interpretation of the test scores, while reliability is related to the consistency of test scores. Both are interrelated because reliability will affect the validity of the measurement, but not everything reliable is valid (Nurhalimah et al., 2022).

The validity of the 50 items developed by the teachers in this study was measured by the researchers' colleagues. They were Indonesian language teachers at SMPN 8 Yogyakarta. The validity of the test instrument according to (Nurgiyantoro, 2016) is shown by several indicators. The item validity was measured by the colleagues using the indicators proposed by the experts. The indicators include some aspects, namely material, construction, and language. First, the material aspect consists of (a) the conformity of the items with indicators in the blueprint, (b) the conformity of the items with science, (c) the answer key, and (d) the distractor function. Second, the construction aspect includes: (a) the clarity of item formulation, (b) the clarity of the options, (c) the homogeneous answer choices, (d) no items with double negative statements, (e) the length of each item, (f) dependency of each item, and (g) the order of items including numbers and time. At last, the language aspect focuses on (a) the communicative language used in the items, (b) grammar, (c) sentences with a double meaning, and (d) the use of vocabulary (Nurgiyantoro et al., 2020)

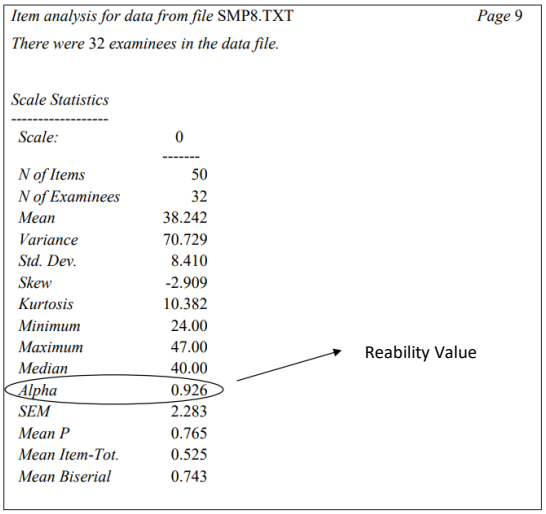
After the 40 items were examined, it can be concluded that all items were categorized as valid items because all of them were considered appropriate. Based on the table above, there were fifty items categorized as valid items (100%). They were item numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, and 50.

Based on the result above, it was found that there was no invalid item (0%), and all items were considered valid (100%). Therefore, it can be stated that all items in the final exam have a high level of validity. Himawan & Nurgiyantoro (2022) pointed out that in testing content validity, teachers may compare the developed items with the blueprint or curriculum that has been taught. This process can be done by the teachers and their peers.

2. Reliability

A test instrument is considered reliable if it generates the same results when used to assess anyone at any time. In terms of reliability, the researchers used ITEMAN 3.0 to find the Alpha

value. The result of the reliability analysis is presented in the ITEMAN output below.



Item analysis for data from file SMP8.TXT *Page* 9

There were 32 *examinees in the data file.*

Scale Statistics

Scale:	0
N of Items	50
N of Examinees	32
Mean	38.242
Variance	70.729
Std. Dev.	8.410
Skew	-2.909
Kurtosis	10.382
Minimum	24.00
Maximum	47.00
Median	40.00
Alpha	0.926
SEM	2.283
Mean P	0.765
Mean Item-Tot.	0.525
Mean Biserial	0.743

Figure 1. Reliability Test Results Through the ITEMAN Program

A test instrument is considered reliable if it generates the same results when it is used to assess anyone at any time. In terms of reliability, the researchers used ITEMAN 3.0 to get the Alpha value. Putri & Ofianto (2019) define reliability as a measure that shows the level of consistency of a test item. In this research, the reliability value obtained was 0.926, indicating that the developed multiple-choice questions have a very high level of reliability. The results are in line with a study conducted by Nurhalimah et al (2022). It is shown that a coefficient that is higher or equal to 0.20 indicates a very high level of reliability (Nayla Amalia & Widayati, 2012; Suharti, 2017; Ida & Musyarofah, 2021)

The results of the reliability test questions are in accordance with research (Fernanda & Hidayah, 2020) which stated that the reliability test of test instruments using a classical test theory approach can be seen from the alpha score obtained. In a test, it is important to observe the consistency and certainty, which are reflected in the test results obtained (Nuryanti et al., 2018). thus making it trustworthy or dependable. The Alpha score in the ITEMAN output can be used as a guide in item analysis, as Alpha is closely related to the reliability of a test instrument

The alpha score can be used as a determinant of the quality of criteria questions with reliability classification: 0.00-0.20 (very low); 0.21-0.40 (low); 0.41-0.70 (medium); 0.71-0.90 (high); 0.91-1.00 (very high). Test reliability refers to understanding whether a test can consistently measure something that will be measured from time to time. Measurement results can be trusted only if relatively similar results are obtained several times on the same group of subjects, as long as the aspect being measured in the subject has not changed. The meaning of reliability of measuring instruments and reliability of measuring results are usually considered the same (Erfan et al., 2020); (Hanifah, 2014); (Kurniawan, 2015).

3. Item Discrimination Index (IDB)

Item discrimination index shows whether an item can distinguish the abilities of the test takers in the high and low groups (Nurgiyantoro, 2016). The Item Discrimination Index (IDB) indicates the extent to which an item can differentiate between the abilities of participants in high and low groups (Nurgiyantoro, 2016). A good test item should have a high and positive coefficient on

the correct answer. The IDB can be seen from the ITEMAN output in the Item Discrimination section. It is stated that an IDB ranging from 0.00 to 0.20 is considered poor, an IDB from 0.21 to 0.40 is deemed adequate, an IDB from 0.41 to 0.70 is considered good, and an IDB from 0.71 to 1.00 is considered very good. Negative results indicate poor discrimination.

The discrimination index of the final exam for the Indonesian language subject for the class VIII students of SMPN 8 Yogyakarta is presented in the table below.

Table 2. IDB Study Results

No	Category	Number of Items	Amount
1	Failed	10, 21, 22	3
2	Enough	24, 29, 38,42	4
3	Medium	11,14, 17, 19, 20	5
4	Good	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 18, 23, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 43, 44, 45, 46, 47, 48, 49, 50	38
5	Very Good	-	0

Item discrimination index shows the ability of an item to distinguish the abilities of the test takers in the high and low groups (Nurgiyantoro, 2016). Test items with a high item discrimination index should have answers with a positive and high coefficient. For the purpose of learning in the classroom, a more moderate strategy can be implemented by considering a discrimination index higher or equal to 0.20 as a qualified item (Nurgiyantoro, 2016)

Based on the results of the item discrimination analysis, good test items have discrimination indices between 0.40 and 1.00. In this study, there are 38 items with Very Good discriminating indices (item numbers 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 18, 23, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 43, 44, 45, 46, 47, 48, 49, 50). Items with good discrimination indices are numbers 11, 14, 17, 19, and 20. Those items have indices between 0.30 and 0.39. Then, items that belong to the Fair category are numbers 24, 29, 38, and 42. The discrimination indices of those items are between 0.20 s.d. 0.29. At last, items having Poor indices (below 0.20) are numbers 10, 21, and 22. Those items were then deleted.

The categorization proposed by (Anggraini & Suyata, 2014) shows that indices between 0.00 and 0.20 belong to the Poor category, indices from 0.21 to 0.40 belong to the Fair category, indices ranging from 0.41 to 0.70 belong to the Good category, the indices between 0.71 and 1.00 belong to the Very Good category, and negative indices indicate Very Poor discriminators (Himawan & Nurgiyantoro, 2022).

4. Item Difficulty Index

The level of difficulty can be seen from the ITEMAN analysis in the item difficulty section. ITK is an index that indicates how easy or difficult a test item is for the tested participants (Nurgiyantoro, 2016). The range of this difficulty index is from 0.00 to 1.00. If the difficulty index shows a value of 1.00, then the item is not difficult to answer. Conversely, if the difficulty index is 0.00, then the item is very difficult to answer. ITK significantly determines the difficulty of the questions being tested, making it closely related to critical thinking skills.

Table 3. ITK Study Results

No	Category	Range	Item Number	Amount
1	Very Easy	0.81-1.00	1, 2, 3, 5, 6, 7, 8, 9, 13, 14, 15, 16, 18, 23, 25, 26, 28, 31, 32, 33, 34, 35, 37, 39, 41, 43, 44, 45, 46, 49, 50	31
2	Easy	0.61-0.80	4, 11, 12, 22, 27, 30, 36, 40, 42, 47	10
3	Currently	0.41-0.60	19, 20, 29, 38, 48	5
4	Difficult	0.21-040	17, 21	2
5	Very Difficult	0.00-0.20	10, 24	2

Item difficulty is an index that represents how easy or difficult the item is according to test takers (Nurgiyantoro, 2016). Item difficulty index ranges from 0.00 to. 1.00. If the index is 1.00, the item is not difficult to answer. Conversely, if the difficulty index is 0.00, the question is very difficult to answer. The item difficulty index ranging from 0.20-0.80 is tolerable (Nurgiyantoro, 2016). There are some categories of difficulty index, namely Very Difficult (0.00-0.199) Difficult (0.20-0.40), Moderately Difficult (0.41-0.60), and Easy (0.61-0.80).

Based on the results of this present study, five items belong to the Moderately Difficult category (item numbers 19, 20, 29, 38, and 48). Ten items are in the Easy category (numbers 4, 11, 12, 22, 27, 30, 36, 40, 42, and 47). There are two Very Difficult questions (numbers 10 and 24). Then, there are 31 Very Easy items (numbers 1, 2, 3, 5, 6, 7, 8, 9, 13, 14, 15, 16, 18, 23, 25, 26, 28, 31, 32, 33, 34, 35, 37, 39, 41, 43, 44, 45, 46, 49, and 50).

The categories of good and poor discrimination significantly impact the quality of classroom learning and, consequently, determine the quality of students in solving problems. Questions can be considered good if their cognitive level distribution is even, thereby it can teach students to solve various conceptual problems through evaluation questions developed during classroom learning (Pujiastuti & Kulup, 2021); (Azizah & Sumardi, 2021).

5. Distractor Function

Distractors are used to identify test takers with high ability. The distractor is said to function effectively if there are more test-takers with low ability selecting it. Meanwhile, if there are more test takers with high ability selecting the distractor, the distractor does not function well (Iskandar & Rizal, 2018). Distractors are effective when they are selected by test takers; however, if any distractors receive no selections, they are considered ineffective (Arvianto, 2016). The function of the distractor in this study is presented as follows. The effectiveness of distractors can be assessed by examining the proportional endorsing values provided in the ITEMAN output. Distractors are considered effective if they are selected by at least 5% of test participants. A question is considered suitable if it includes effective distractors that surpass the 5% threshold. Distractors are derived from analyzing test takers' responses to incorrect answer options (Syarifah et al., 2020).

Table 4. Results of Study of Distractor Items

No	Category	Item	Amount
1	Works	2, 3, 10, 11, 12, 18, 19, 22, 27, 30, 33, 37, 38, 40, 42, 47, 48, 49	18
2	Not Working	1, 4, 5, 6, 7, 8, 9, 13, 14, 15, 16, 17, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 34, 35, 36, 37, 39, 41, 43, 44, 45, 46, 50	32

Nurgiyantoro (2016) states that there are several criteria for determining the effectiveness of a distractor, namely: (1) all distractors (wrong options) must be selected, (2)

test takers from a high group should select less number of wrong options than the test takers from low-group, and (3) if there is only one test-taker chooses one of the wrong options, he belongs to the low-group. Criteria 2 and 3 are often seen as burdensome since they are similar to the concepts of Distractor Function. Thus, criterion 1 is more effectively used. Wrong options are not effective if no test taker chooses them. This type of option should be deleted or revised (Mardiana & Suyata, 2017).

Based on the results of distractor function analysis, it was found that there are effective and ineffective distractors. A set of distractors is considered effective when all or three of the options work well, and a distractor is considered ineffective if one, two, or three distractors do not work. There are 18 items with effective distractors (numbers 2, 3, 10, 11, 12, 18, 19, 22, 27, 30, 33, 37, 38, 40, 42, 47, 48, and 49). Meanwhile, the other 32 items have ineffective distractors (numbers 1, 4, 5, 6, 7, 8, 9, 13, 14, 15, 16, 17, 20, 21, 23, 24, 25, 26, 28, 29, 31, 32, 34, 35, 36, 39, 41, 43, 44, 45, 46, 50). According to (Putri & Ofianto, 2019) effective distractors are those chosen evenly by test takers. Meanwhile, ineffective test items are unequally chosen by the takers.

4. Conclusion

Based on the results of an analysis of the quality of the end-of-semester assessment questions for the Indonesian language subject at SMPN 8 Yogyakarta class VIII. Can be explained as follows.

Validity was carried out to colleagues, by analyzing 50 questions, and giving colleagues a validity questionnaire containing indicators of the validity of the questions, according to the expert. Based on the results of the test reliability analysis, an Alpha result of 0.926 was obtained. This shows that the reliability of the multiple-choice questions in PAS Odd Indonesian Class VIII-B at SMP Negeri 8 Yogyakarta is categorized as very high.

Based on the results of the IDB analysis, it was found that the categories of items that had good discriminating power were 0.40 to.d. 1.00. In this study, it was found that there were 38 items with good discriminating power, Item categories which has a moderate difference power is 0.30 to.d. 0.39. There are 5 average item power. The categories of items that have sufficient discriminating power are 0.20 to.d. 0.29. There are only 4 items in power. The categories of items that have poor discriminatory power or fail are below 0.20. There are 3 bad or failed item items. Based on the results of the study, there were 5 difficulty levels of students with moderate item categories. There were 10 difficulty levels of students with easy item categories. There are 2 difficulty levels of students with very difficult item categories. There are 2 levels of difficulty for students with very easy item categories 31 (Yusmilda et al., 2023).

Based on the results of an analysis of the functioning of the distractor based on the results of the multiple-choice item items in PAS Odd Indonesian Language Class VIII at SMP Negeri 8 Yogyakarta, it is known that a total of 50 items have a functioning distractor and a non-functioning distractor. The item distractors work if all or 3 of the distractors work. A distractor doesn't work if you have 1, 2, or 3 distractors that don't work. The results of the analysis of the distractor items, in the developed questions, show that there are 18 items with good distractor function, and there are 32 questions with distractors that don't work properly. Overall, the results of this analysis aim to provide a reference point, teachers. Dealing with item analysis, especially in Indonesian language subjects, which is currently still rarely done.

In summary, the implications of the research findings can be used as a reference for teachers in presenting evaluation instruments,

including grids, assessment guidelines, and questions, thus determining the quality of the items tested on students. In relation to this, the process of higher-order thinking skills (HOTS) can be easily applied by teachers to students. Aspects of reading skills in learning, especially in Indonesian language-based text learning, will assist teachers and students in implementing reading literacy in teaching.

This research enables teachers to conduct evaluations through steps such as: (1) drafting question grids, (2) conducting question validity tests with experts or peers, (3) revising and analyzing evaluation questions using the ITEMAN program, and (4) examining the ITEMAN output to construct genuinely effective evaluation instruments, considering aspects like reliability, IDB, ITK, and the functionality of distractors.

5. References

Al-faruq, Z. (2023). Peran Penggunaan Desain Evaluasi Untuk Meningkatkan Kualitas Pembelajaran. *Ilma Jurnal Pendidikan Islam*, *1*(2), 158–171. https://doi.org/10.58569/ilma.v1i2.587

Alfarisa, F., Chudari, I. N., & Robiansyah, F. (2019). Analisis Butir Soal IPS Kelas V Sekolah Dasar Menggunakan Software ITEMAN. *EduBasic Journal: Jurnal Pendidikan Dasar*, *1*(2), 100–106. https://doi.org/10.17509/ebj.v1i2.26474

Alnovgada, V. R. S., & Suyata, P. (2019). *The Effectiveness of Picture and picture Cooperative Learning Models of Writing Instructions Skills in Class VIII Students of SMP Negeri 2 Sui Ambawang*. *297*(Icille 2018), 432–437. https://doi.org/10.2991/icille-18.2019.90

Anggraini, D., & Suyata, P. (2014). Karakteristik Soal Uasbn Mata Pelajaran Bahasa Indonesia Di Daerah Istimewa Yogyakarta Pada Tahun Pelajaran 2008/2009. *Jurnal Prima Edukasia*, *2*(1), 57. https://doi.org/10.21831/jpe.v2i1.2644

Arifin, Z., & Retnawati, H. (2017). Pengembangan instrumen pengukur higher order thinking skills matematika siswa SMA kelas X. *PYTHAGORAS: Jurnal Pendidikan Matematika*, *12*(1), 98. https://doi.org/10.21831/pg.v12i1.14058

Arvianto, I. R. (2016). Pemanfaatan program Iteman 3.0 untuk analisis butir soal lomba cerdas cermat teknologi informasi dan komunikasi tingkat SMA sederajat. *Jurnal Teknologi Informasi*, *XI*(33), 1–13.

Asrial, A., Syahrial, S., Sabil, H., Kurniawan, D. A., Perdana, R., Nawahdani, A. M., Widodi, B., Rahmi, R., & Nyirahabimana, P. (2023). Quantitative Analysis Of Elementary School Students' Curiosity and Web-Based Assessment Responses. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *5*(2), 107–119. https://doi.org/10.23917/ijolae.v5i2.21646

Azizah, N., & Sumardi, H. (2021). Analisis Kualitas Dan Tingkat Kognitif Soal Matematika Penilaian Akhir Semester (Pas) Ganjil Kelas Ix Di Smp N 10 Kota Bengkulu Tahun 2020/2021. *Journal Mathematics Education Sigma [JMES]*, *2*(2). https://doi.org/10.30596/jmes.v2i2.7936

Dewi, S. K., & Sudaryanto, A. (2020). Validitas dan Reliabilitas Kuesioner Pengetahuan , Sikap dan Perilaku Pencegahan Demam Berdarah. *Seminar Nasional Keperawatan Universitas Muhammadiyah Surakarta (SEMNASKEP) 2020*, 73–79.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Tes Klasik Dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19.

Fernanda, J. W., & Hidayah, N. (2020). Analisis Kualitas Soal Ujian Statistika Menggunakan Classical Test Theory dan Rasch Model. *Square : Journal of Mathematics and Mathematics Education*, *2*(1), 49.

https://doi.org/10.21580/square.2020.2.1.5363

Fitriani, F., Ibrahim, I., & Nugroho, E. D. (2020). Analisis Soal Ujian Akhir Semester Pada Mata Pelajaran Ipa Berdasarkan Dimensi Proses Kognitif Taksonomi Anderson Dankemampuan Berpikir Kritisdi Smp Negeri 1 Nunukan Selatan. *Biopedagogia*, *2*(1), 37–43. https://doi.org/10.35334/biopedagogia.v2i1.1716

Fridaram, O., Istharini, E., Cicilia, P. G. C., Nuryani, A., & Wibowo, D. H. (2021). Meningkatkan Konsentrasi Belajar Peserta Didik dengan Bimbingan Klasikal Metode Cooperative Learning Tipe Jigsaw. *Magistrorum et Scholarium: Jurnal Pengabdian Masyarakat*, *1*(2), 161–170. https://doi.org/10.24246/jms.v1i22020p161-170

Hanifah, N. (2014). Perbandingan Tingkat Kesukaran, Daya Pembeda Butir Soal Dan Reliabilitas Tes Bentuk Pilihan Ganda Biasa Dan Pilihan Ganda Asosiasi Mata Pelajaran Ekonomi. *SOSIO E-KONS*, *6*(1), 41–55.

Himawan, Suyata, K. (2024). Developing Project-Based Learning-Based eBook "Critical and Creative Reading" to Improve Students' Critical Thinking Skills Riswanda. *Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran*, *10*(1), 392–404.

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 (Analysis of exercise items for odd semester end of semester Indonesian language subjects class. *Kembara: Jurnal Keilmuan Bahasa, Sastra, Dan Pengajarannya*, *8*(1), 160–180.

Himawan, R., & Suyata, P. (2022). DEVELOPING HOTS QUESTIONS : EVALUATING PERSUASIVE SPEECH TEXT LEARNING IN GRADE IX OF JUNIOR HIGH SCHOOL. *Jurnal Gramatika: Jurnal Penelitian Pendidikan Bahasa Dan Sastra Indonesia*, *8*(1), 50–64.

Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'Arrib: Journal of Arabic Education*, *1*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Kurniawan, T. (2015). Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran IPS Sekolah Dasar (Analysis of Odd Semester Final Test Items in Elementary School of Social Studies Subjects). *Journal of Elementary Education*, *4*(1), 1–6.

Kusumaningtyas, D. A., Manyunu, M., Kurniasari, E., Awalin, A. N., Rahmaniati, R., & Febriyanti, A. (2024). Enhancing Learning Outcomes: A Study on the Development of Higher Order Thinking Skills based Evaluation Instruments for Work and Energy in High School Physics. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *6*(1), 14–31. https://doi.org/10.23917/ijolae.v6i1.23125

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan. *BINTANG : Jurnal Pendidikan Dan Sains*, *3*(2), 198–214.

Mardiana, M., & Suyata, P. (2017). Evaluating the philosophical foundation of 2013 Curriculum. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *21*(2), 175–188. https://doi.org/10.21831/pep.v21i2.13336

Martin, T. I. H. (2020). Pengembangan Instrumen Soal HOTS (High Order Thinking Skill) Pada Mata Kuliah Fisika Dasar 1. *Jurnal Pendidikan Fisika*, *8*(1), 18–21.

Muhith, A. (2018). Problematika Pembelajaran Tematik Terpadu di Min III Bondowoso. *Indonesian Journal of Islamic Teaching*, *1*(1), 45–61.

Mustafidah, H., Hartati, S., Wardoyo, R., & Suyata, P. (2021). Intelligent computational model to determine the order of thinking skills of test items. *ICIC Express Letters*, *15*(9), 999–1006. https://doi.org/10.24507/icicel.15.09.999

Nanda Pratiwiningtyas, B., Susilaningsih, E., & Made Sudana, I. (2017). Pengembangan Instrumen Penilaian Kognitif untuk Mengukur Literasi Membaca Bahasa Indonesia Berbasis Model Pirls pada Siswa Kelas IV SD. *Journal of Educational Research and Evaluation Sejarah Artikel*, *6*(1), 1–9.

Nayla Amalia, A., & Widayati, A. (2012). Analisis Butir Soal Tes Kendali Mutu Kelas XII Sma Mata Pelajaran Ekonomi Akuntansi Di Kota Yogyakarta. *Jurnal Pendidikan Akuntansi Indonesia*, *X*(1), 1–26.

Nurgiyantoro, B. (2016). *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*. BPFE-Yogyakarta.

Nurgiyantoro, B., Lestyarini, B., & Rahayu, D. H. (2020). Konstruk Asesmen Literasi Fungsional Untuk Siswa Sekolah Menengah Pertama. *Litera*, *19*(2), 194–211. https://doi.org/10.21831/ltr.v19i2.32977

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan Antara Validitas Item Dengan Daya Pembeda Dan Tingkat Kesukaran Soal Pilihan Ganda Pas. *Natural Science Education Research*, *4*(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Analisis Iteman dan Model Rasch pada Pengembangan Instrumen Kemampuan Berpikir Kritis Peserta Didik Sekolah Menengah Kejuruan. *Jurnal Inovasi Pendidikan IPA*, *4*(2), 224–233.

Pangesti, F., Fauzan, F., & Risnawati, R. (2020). Kualitas butir soal try out uji pengetahuan dalam memprediksi tingkat kelulusan mahasiswa PPG. *Jurnal Pendidikan Profesi …*, *1*(2), 91–98.

Parancika, R. B., & Suyata, P. (2020). Implementasi Pembelajaran Menulis Teks Eksplanasi Kompleks Pada Siswa Kelas XI SMAN 10 Yogyakarta Dengan Menggunakan Strategi Writing a Story Based on a Picture / Photograph. *Rumpun Jurnal Persatuan Melayu*, *8*(1), 13–25.

Prastikawati, E. F., Adeoye, M. A., & Ryan, J. C. (2024). Fostering Effective Teaching Practices: Integrating Formative Assessment and Mentorship in Indonesian Preservice Teacher Education. *Indonesian Journal on Learning and Advanced Education*, *6*(2), 230–253. https://doi.org/10.23917/ijolae.v6i2.23431

Pujiastuti, R., & Kulup, L. I. (2021). Penyusunan Instrumen Penilaian Kognitif Berbasis HOTS Melalui Problem Based Learning dan Peer Assessment. *Indonesian Language Education and Literature*, *7*(1), 88. https://doi.org/10.24235/ileal.v7i1.9058

Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMA N 1 Kutowinangun Tahun Pelajaran 2019/2022 Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, *10*(2), 205–214.

Putri Pangestu, D., & Rohinah, R. (2019). Pengaruh Kesiapan Belajar Terhadap Keaktifan Peserta Didik dalam Proses Pembelajaran AUD. *Golden Age: Jurnal Ilmiah Tumbuh Kembang Anak Usia Dini*, *3*(2), 81–90. https://doi.org/10.14421/jga.2018.32-02

Putri, R. H., & Ofianto. (2019). Efektivitas Analisis Butir Menggunakan Anajohn, Anates dan Iteman Studi Soal USBN

Pelajaran Sejarah Kota Padang. *Jurnal Mahasiswa Ilmu Sejarah Dan Pendidikan*, *1*(2), 1–11.

Retnawati, H. (2015). *Validitas, Realibilitas & Karakteristik Butir*. Parama Publishing.

Rotama, A. D., Budiutomo, T. W., & Bowo, A. N. A. (2020). Analisis Butir Soal Penilaian Tengah Semester Mata Pelajaran PPKn Kelas VII di Smp Muhammadiyah 7 Yogyakarta. *Academy of Education Journal*, *11*(01), 24–35. https://doi.org/10.47200/aoej.v11i01.314

Ruay Garcés, R. (2018). La Evaluación: Una estrategia para desarrollar Aprendizajes Profundos en el estudiante. *Boletín Redipe*, *7*, 47–62.

Setiawan, K. E. P., Yudha, R. K., & Arwansyah, Y. B. (2022). Analisis Butir Soal Penilaian Akhir Tahun (PAT) Bahasa Indonesia Kelas XI SMA Negeri 1 Polanharjo Klaten. *Lingua Rima*, *11*(2), 25–33.

Setiawan, M. A., Susongko, P., & Hayati, M. N. (2020). Pendeteksian DIF pada Perangkat Tes Objektif Penilaian Akhir Semester IPA dengan Mengunakan Permodelan Rasch. *Pancasakti Science Education Journal*, *5*(2), 23–29.

Shanta Monica, Y. sudarman. (2013). Analisis Butir Soal Ujian Tengah Semester Ganjil Seni Budaya Kelas VII Di SMPN 29 Sijunjung. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Suharti, S. (2017). Kualitas Tes Bahasa Arab dan Prestasi Peserta Didik Madrasah Tsanawiyah Kabupaten Bantul (Analisis Butir Soal UAMBN. *Jurnal Pendidikan Madrasah*, *2*(1), 185–196.

Susanto, H., Rinaldi, A., & Novalia. (2015). Analisis Validitas Reabilitas Tingkat Kesukaran dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. *Al-Jabar: Jurnal Pendidikan Matematika*, *6*(2), 343.

Syaifudin. (2020). Validitas dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab. *Jurnal Kajian Perbatasa Antarnegara*, *3*(2), 106–118.

Syarifah, L. L., Yenni, Y., & Dewi, W. K. (2020). Analisis Soal-Soal Pada Buku Ajar Matematika Siswa Kelas XI Ditinjau Dari Aspek Kognitif. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *4*(2), 1259–1272. https://doi.org/10.31004/cendekia.v4i2.335

Timor, U., Bahasa, P., Km, J., & Sasi, K. (2022). *Pengembangan Soal Literasi Membaca Model Pisa Development of Pisa Model Reading Literacy Questions Based on*. *7*, 42–50.

Wahyuni, A., & Kurniawan, P. (2018). Hubungan Kemampuan Berpikir Kreatif Terhadap Hasil Belajar Mahasiswa. *Matematika*, *17*(2), 1–8. https://doi.org/10.29313/jmtm.v17i2.4114

Wijaya, A., Eresti, A., Despa, D., & Walid, A. (2019). Analisis Butir Soal Persiapan Ujian Nasional IPA SMP/MTS Tahun 2018 Sampai Dengan 2019 Berdasarkan Taksonomi Bloom. *LENSA (Lentera Sains): Jurnal Pendidikan IPA*, *9*(2), 57–63. https://doi.org/10.24929/lensa.v9i2.78

Yadi, H. (2017). Validitas isi: tahap awal pengembangan kuesioner. *Jurnal Riset Manajemen Dan Bisnis (JRMB) Fakultas Ekonomi UNIAT*, *2*(2), 169–178.

Yusmilda, Y., Budi, I. S., & Zuhad, H. (2023). Pengembangan Instrumen Penilaian Tes Berbasis HOTS Pada Jenjang Pendidikan Dasar Di Era Society 5.0. *Al-Madrasah: Jurnal Pendidikan Madrasah Ibtidaiyah*, *7*(1), 429. https://doi.org/10.35931/am.v7i1.1885