Jurnal Penelitian dan Evaluasi Pendidikan
Volume 25, No 1, June 2021 (108-117)
Online: http://journal.uny.ac.id/index.php/jpep

Analysis of mathematics test items quality for high school

Budi Manfaat; Ayu Nurazizah*; Muhamad Ali Misri
Institut Agama Islam Negeri Syekh Nurjati Cirebon
Jl. Perjuangan, Sunyaragi, Kesambi, Kota Cirebon, Jawa Barat 45132, Indonesia
*Corresponding Author. E-mail: ayunurazizah@mail.syekhnurjati.ac.id

ARTICLE INFO

Article History
Submitted:
4 March 2021
Revised:
14 July 2021
Accepted:
1 August 2021

Keywords
item analysis; quality test; mathematics test

Scan Me:

ABSTRACT

This study aims to determine the quality of the mathematics test items at high school in terms of validity, reliability, differentiation, difficulty level, and distractor effectiveness. This study is an evaluation type of research with a quantitative approach. The subjects in this study were 44 class XII students of SMKN 3 Kuningan and 39 class XII students of SMAN 1 Jalaksana. The results show that the majority (96.67%) of the items are declared valid in content by the experts. The test has very high reliability (0.90). The items have the ideal difficulty level. Most of the questions (70%) have medium difficulty, a few questions (6.67%) are very easy, and a few questions (20%) are difficult, and (3.3%) are very difficult. Most of the items (83.33%) have good discriminating power, and only a few questions (16.67%) have poor discriminating power. Most (90%) of the questions have a well-functioning answer choice, and only a few questions (10%) have the answer choice not functioning properly. Overall, this study can be concluded that the Mathematics Test Questions at SMKN 3 Kuningan are of good quality.

How to cite:
Manfaat, B., Nurazizah, A., & Misri, M. (2021). Analysis of mathematics test items quality for high school. *Jurnal Penelitian dan Evaluasi Pendidikan, 25*(1), 108-117. doi:https://doi.org/10.21831/pep.v25i1.39174

INTRODUCTION

Assessment is one of the five main tasks of teachers in learning, done by collecting and processing information to measure the achievement of student learning outcomes. Gronlund and Linn interpret assessment as a systematic process that collects, analyzes, and interprets information to decide how far the student or a group of students achieve their learning goals either in knowledge aspect, attitude, or creativity (Kusaeri & Suprananto, 2012). Assessment can also be interpreted as collecting proof using some instruments, which can also be used to decide whether the assessor can be recommended as competent or not (Setiawan, 2018). Thus, it can be concluded that assessment is a process of collecting information that can be done after the learning process to increase the student's achievement in the learning process.

The activity of assessing learning needs to be done to measure the level of success of a student as a student. The goal of assessment is to know student's will of learning in cognitive, affective, and psychomotor (Sukanti, 2011). Another goal of assessment is getting information about the student's learning process and the goals of teaching itself (Ratnawulan & Rusdiana, 2015). Therefore, the purpose of the implementation of the assessment of student learning outcomes is to measure students' success in mastering predetermined competencies and measuring the extent of the level of the student's understanding of the learning material delivered by the teacher. Thus, each teacher must assess the mastery of knowledge possessed by students during the learning phase in school. Besides, teachers must participate in the creation of evaluation methods for student learning outcomes and the analysis of student learning outcomes. One of the tools for assessing teachers as educators is their ability to assess learning.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

Given the importance of placing assessment in learning, teachers as evaluators are required to be able to understand and make assessment tools. One of the assessment tools in the learning process is a test. Testing is a technique widely used in education. Goodenough (Baskoro, 2018) defines a test as a task or a series of tasks given to an individual or group of individuals with the intention of comparing their skills with each other. One of the assessment methods used to assess student learning outcomes is the exam. Students should be given tasks as part of the test method to monitor and measure their actions.

It can be considered a good test if the tests are valid and reliable. This means that in terms of preparation, it has met the rules of writing questions, both in terms of construction, substance, or material aspects. Each item should have content validity. This means that the evaluation tool actually contains the material to be measured so that the conformity between the evaluation tool and the content of the material that should be measured is actually displayed in the preparation of the questions. In addition, an assessment tool has high reliability if it shows the accuracy of the results. In other words, the person being tested will get almost the same score if he is tested again with the same test equipment.

In order to measure students' actual abilities, the test kites to measure students' learning outcomes must meet high-quality test qualifications. A higher-quality test instrument can capture students' abilities as individuals or as a group and provide accurate data. This can assist teachers in improving the quality of their students' learning.

To obtain a quality test, a teacher needs to carry out an item analysis. Item analysis is one of the important activities in test development to get quality questions. The goal is to examine each item in order to obtain quality questions before use. In addition, the objective of question analysis is also useful for improving the quality of the items by revising the questions and removing ineffective questions (Kusaeri, 2014). Thus, assessing student learning outcomes or the test can describe a sample of behavior and produce objective and accurate scores.

There are three prerequisites for the assessment or test instrument to be valid (Kusaeri & Suprananto, 2012). A test is said to be valid if it can measure what you want to measure. The consistency of measurement results is referred to as reliability. This means that if a student's answer to the test is tested repeatedly, it will produce consistent results, whereas the usable is related to the practical meaning of the test procedure.

In assessing learning outcomes, the tests are expected to describe behavioral samples, produce objective and accurate scores (Iskandar & Rizal, 2018). According to Suryabrata, a good test is must be standardized, objective, and discriminatory (Suharman, 2018). The test must be standardized, which means that each student must receive the same treatment in terms of test material, scoring, and interpretation of test results so that students who receive a certain score in one location will receive the same score for students on the spot. The test must be objective, which means that one teacher will assess the same test-taker as another teacher. Furthermore, the test must be discriminatory, which means it must be able to detect differences in a symptom found in each student.

The test is said to be a good question if it meets the requirements such as eligibility, level of difficulty, distinguishing power, pattern distribution of answers, and the relationship or correlation of each item with the overall score (Wijayanti, 2020). To find these things, an assessment activity is needed through item analysis which aims to obtain information about questions that have met the requirements of good questions.

Not all tests used are good tests. Research on item analysis has been carried out; for example, Maryanes et al. (2018) showed the results of research on the analysis of the quality of the Mathematics final examination items, namely most (80%) of the items were declared invalid. The reliability coefficient ranges from 0.0824 < 0.70. This question means that it includes questions that have very low reliability. The difficulty level obtained results 37% in the medium category and 63% for the difficult category. The difference power got 67% bad question category, 10% good question category, 23% enough question category. This is also supported by Hamimi et al. (2020), who analyzed the Mathematics Exam items for Class VII Odd Semester in the 2017/2018 academic year. The result of the research was that most of the questions used were invalid because there were questions that had low validity, and some were even very low. In addition, this question also has a low degree of reliability or can be said to be unreliable. The results of research conducted by Tilaar and Hasriyanti (2019) show that (1) for the type of multiple-choice questions, few questions (16.67%) had very good quality so that the questions could be saved for reuse. As many as 50.00% of the questions still need to be revised, and 33.33% have very bad quality, so that they cannot be stored in the question bank. (2) For the type of description questions, 40.00% of the items had good quality, 40.00% of the items needed revision, and 20.00% of the items had poor quality. This study indicates that the majority of the questions did not meet the criteria for good questions and must be revised.

Based on interviews with several mathematics teachers at SMKN 3 Kuningan, one of the favorite schools in Kuningan Regency, teachers usually took questions from various mathematics books in compiling tests, both for practice and final semester assessments. However, the results obtained from the test are still many students who get scores below the SKM (*Skor Ketuntasan Minimal*/Minimum Completeness Score). In fact, the teachers at the school already have good educational qualifications, and the school has adequate facilities and infrastructure.

It means that there is a possibility that the student's learning outcomes do not describe the actual situation regarding students' abilities. A learning result is obtained using a test that does not describe a student's learning achievement. It may be the score of each student that higher or lower than the actual ability. The results of this study will provide erroneous information about the achievement of learning outcomes. Therefore, the quality of the test is one of the factors that need to be addressed in learning evaluation activities. In addition, many teachers still do not pay attention to the quality of the items because they have not analyzed the items, so the items used cannot detect whether the questions have met the criteria or not.

From the aforementioned statement, it can be seen that teachers pay less attention to the principles in assessing student learning outcomes (Zuhera et al., 2017), causing the learning outcomes to be out of sync with the students' abilities. This will be risky if the assessment of learning outcomes is not in accordance with students' abilities. Thus, the researchers wanted to know the quality of the mathematics test items used at SMKN 3 Kuningan.

RESEARCH METHOD

This study is an evaluation research. In this study, the researchers evaluate the items contained in the Mathematics Module Class XI book SMKN 3 Kuningan, which aims to determine the quality of the items in terms of validity, reliability, difficulty level, discriminating power, and deceptive effectiveness. The subjects in this study were 44 students of class XII of SMKN 3 Kuningan and 39 students of class XII of SMAN 1 Jalaksana. The average score for the math test at SMKN 3 Kuningan is 32.3, and the average score for the math test at SMAN 1 Jalaksana is 48.3. Thus, the overall math test average score is 39.8.

In this study, the target population was all items of mathematics module for class XI SMKN 3 Kuningan. The available population was all of the items in the mathematics module for class XI published in 2019. The sample consisted of 30 multiple-choice questions drawn from each chapter. In choosing 30 items, the researchers adjusted it to the test time of 120 minutes.

The data obtained from this study are quantitative data and data from instrument validation by an expert or expert judgment. The quantitative data in question are data obtained from students' responses or answers to the questions given to be analyzed in terms of quality. The sources of data in this study are validators or experts (expert judgment) and students. To assess whether the test instrument was valid or not, the researchers asked three validators or experts to assess it.

https://doi.org/10.21831/pep.v25i1.39174

The technique used by researchers in collecting data is the documentation technique. The researchers collected documents in the form of test sheets taken by the mathematics module and student answer sheets, and the validation results from the expert judgment (expert judgment). This documentation aims to determine the quality of the items contained in mathematics textbooks.

$$CVR = \frac{ne-\frac{N}{2}}{\frac{N}{2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$r_{tt} = \frac{2\, r_{\frac{11}{22}}}{1+\left| r_{\frac{11}{22}} \right|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\text{Difficulty level} = \frac{\text{numbers of students who answer the item correctly}}{\text{total number of students who answer the item}} \ldots\ldots (3)$$

$$DP = \frac{RU-RL}{\frac{1}{2}T} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

$$IP = \frac{P}{(N-B)(n-1)} \times 100\% \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Content validation analysis uses the CVR (Content Validate Ratio) formula, namely Formula (1), as suggested by Lawshe (Azwar, 2019). Meanwhile, the reliability analysis uses the Spearman-Brown Formula (2) (Azwar, 2019). The difficulty level of the items will be analyzed using Formula (3) (Jannah et al., 2021). The item difference was analyzed using Formula (4) (Jannah et al., 2021). The effectiveness of the trickster was analyzed using the Formula (5) (Rahayu & Djazari, 2016).

FINDINGS AND DISCUSSION

The results of this study were obtained from students' responses and analyzed using the Anates program, which included reliability, difficulty level, discriminating power, and the effectiveness of questioners. The expert's assessment was analyzed for validity in order to obtain a content-valid test. The results and discussion are elaborated as follows.

Validity

Validity is a concept that measures the extent to which the test can measure what should be measured. The test is valid if it can reveal data from the variables appropriately and does not deviate from the actual situation.

Content validity is defined as the degree to which a test's content is reasonable or relevant, as determined by a qualified panel's objective analysis or expert judgment (Azwar, 2019). This validity refers to the degree to which the test has proof of the content's validity as determined by logical analysis or the test's logic. The query indicator was used to change each item's content validity. After the instrument in the form of a test is assessed by experts or expert judgment and calculated using the CVR formula, the results of the content validity are shown in Table 1.

The data in Table 1 show the validation results by the expert that 3.33% (1 item) are declared as invalid questions, and 96.67% (29 items) are stated as valid questions with a CVR value of 1. When compared with the results of the research by Maryanes et al. (2018), Hamimi et al. (2020), and Tilaar and Hasriyanti (2019), the results of the research on the mathematics test items are better because most of the items (96.67%) are valid.

https://doi.org/10.21831/pep.v25i1.39174

Table 1. Result Content Validity Analysis

Category	Number of Items	Number of Questions	Percentage
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	29	96.67%
Invalid	30	1	3.33%

Thus, it can be concluded that the mathematics test questions at SMKN 3 Kuningan are of good quality because most (96.67%) are declared contentedly valid by experts. This means that most of the items are in accordance with the competency indicators to be measured. In other words, the test is structured based on the content of the material being evaluated. The question points contain relevant content and do not go beyond the boundaries of the learning objectives. As for the follow-up to the results of the analysis of the mathematics test items taken from the Mathematics book of Class XI SMKN 3 Kuningan, namely valid items, they are entered into the question bank and can be reused for learning outcomes tests, while invalid items are declared as failed questions should be thrown away.

Reliability

From the analysis of the Anates program, the reliability coefficient of the test is 0.90. According to Fraenkel et al., if an instrument can be said reliable, that is, when the Spearman-Brown reliability coefficient is more than 0.70 (ri > 0.70) (Yusup, 2018). This means that the test results meet a high level of consistency, so the assessment results of the same test-taker using the test repeatedly will show the same results. This means that this mathematics test is of good quality when viewed from the point of view of the reliability of the questions.

One of the characteristics of the high-reliability questions is that the test used consists of the number of valid questions as the result of the validity analysis. Arifin (2013) states that a reliable test is if it has a high coefficient and low standard error of measurement. Problems with very high reliability mean that the test can show consistent results if tested repeatedly on a test taker as long as the aspects measured in the test taker have not changed. Compared to the results of previous studies that obtained low reliability, this math test has a better degree of consistency so that the test is consistent and reliable to use as a measuring tool. When compared to the results of previous studies, Maryanes et al. (2018) obtained very low reliability, while Hamimi et al. (2020) obtained a low or unreliable degree of reliability, this mathematics test has a better level of constancy so that the test is consistent and reliable to be used as a measuring tool.

Level of Difficulty

The level of difficulty is related to how many or at least students answer the questions correctly. The difficulty of a question in a test is seen from the ability of students to answer the question, not from the assumptions of the teacher who compiled the questions. Because questions that are difficult or easy for teachers are not necessarily difficult or easy for students (Wijayanti, 2020). In order to find out the difficulty level of an issue, it is necessary to calculate the difficulty index. The difficulty index ranges from 0.00 to 1.00 (Zein & Darto, 2012). The smaller the difficulty index obtained, the more difficult the question is for students to answer, and vice versa. In principle, the average score obtained by the testee on the item in question is called the item difficulty level (Sumiati et al., 2018). The description of the difficulty level analysis results using the Anates version 4.0.9 program is presented in Table 2.

Table 2 shows that most of the questions (70%) had medium difficulty, a few questions (6.67%) are very easy, and a few questions (20%) are difficult, and (3.3%) are very difficult. A good question is one that is not too easy or not too difficult (Arikunto, 2013). Problems that are too easy will not stimulate students to solve a more difficult problem, and questions that are too difficult will result in a loss of enthusiasm when working on the problem because the questions presented are beyond their ability.

Table 2. Results of Difficulty Level Analysis

Difficulty Category	Number of Items	Number of Questions	Percentage
Very easy	10, 22	2	6.67%
Easy	-	-	0%
Medium	1, 3, 4, 5, 6, 7, 9, 11, 13, 14, 17, 18, 19, 20, 21, 25, 26, 27, 28, 29, 30	21	70%
Difficult	2, 8, 12, 15, 16, 23	6	20%
Very difficult	24	1	3.33%

Compared to the research results of Maryanes et al. (2018), which obtained 67% difficult questions, the results of Tilaar and Hasriyanti (2019) show that 46.67% of the questions are categorized as difficult, so this math test shows an enhancement. Compared to the results of previous studies that obtained 67% of difficult questions, this math test shows an increase. Most of the items used in this study had an ideal difficulty level, namely in the medium category. It can be concluded that the questions contained in this math test are of good quality.

Questions that have a moderate level of difficulty can be stored in the question bank for reuse in other tests. As for the easy and difficult items, three possible follow-ups can be done: items are discarded and not reused in the next test; researching or tracing back what causes the items are so easy to answer or too difficult to answer by students. Then improvements were made so that the items could be reused. Easy items can be reused in loose tests with the aim that most test-takers pass the test. In contrast, difficult items can be reused in very strict tests, usually selection tests, so that they can be stored in a bank as a separate question.

Discriminating Power

Discriminating power, namely the level of ability of the items, can distinguish between smart students (students who have mastered the material) and less intelligent students (students who have not mastered the material). Like the degree of difficulty, the discriminating power also has an index called the discriminating power index (DP). To find the discriminating power of a problem, it is necessary to calculate the index of discriminating power. The discriminatory index ranges from 0.00 to 1.00 (Supriadi et al., 2019). The higher the discriminant index of an item, the better it is at distinguishing between intelligent and less intelligent students (Kusumawati & Hadi, 2018). Based on the analysis of discriminating power using the Anates program, the research results were obtained, as presented in Table 3.

Table 3. Results of Discriminatory Power Analysis

Category Discriminating Power	Number of Items	Number of Questions	Percentage
Very good	1, 3, 6, 9, 13, 18, 19, 25, 26, 27, 28, 29	12	40%
Good	2, 4, 5, 7, 11, 12, 14, 16, 17, 20, 21, 22, 30	13	43.33%
Enough	8, 15, 23	3	10%
Bad	10, 24	2	6.67%
Very bad	-	0	0%

Table 3 shows that most (83.33%) items have good discriminating power. Only a few questions (16.67%) were poorly differentiated. That is, most of the questions can show the difference between high-ability students and low-ability students. In this case, smart students

https://doi.org/10.21831/pep.v25i1.39174

or students in the upper group answered more correctly, while less smart students or students in the lower group mostly could not answer the questions correctly.

When compared to the results of the research by Maryanes et al. (2018), which obtain 67% of the distinguishing power in the poor/bad category, 10% in the good category, 23% in the sufficient category, the research results of Tilaar and Hasriyanti (2019) show that 10% results are in a bad category (negative), 26.67% in the poor category, so this math test shows an increase. Most of the items used in this study have the ability to differentiate between students who are smart and less intelligent students. It can be concluded that the questions contained in this math test are of good quality.

Problems with excellent, good, and sufficient discriminating power can be put in the question bank for reuse. However, questions that are categorized as sufficient if they want to be reused must be revised first. Meanwhile, questions that have bad or bad distinctions should be discarded or revisited. There may be several reasons that allow the item not to show differences in student abilities, including having two or more correct answers, not having the correct answer, the measured competency is not clear, the bully does not function, the material asking questions is too difficult so that many students answer the original -casual, or most students are good at misconceptions or misinformation in understanding questions. Therefore, questions with poor discriminatory power should be reviewed and corrected for reuse in other tests.

Deceptive Effectiveness

In the analysis of multiple-choice items, the effectiveness of distractors is one of the requirements for an item to be said to be of good quality. A good item is the one which the deceiver is chosen evenly by all test takers who answer wrongly (Arifin, 2013). A good distractor is chosen by 5% of the total number of test-takers (Arikunto, 2013) and good distractors affect the result of the test to differentiate the upper group students from the lower ones (Hartati & Yogi, 2019). On the other hand, bad items are those in which the test taker unevenly selects the cheater. Based on the results of the decoy effectiveness analysis, the data in Table 4 are obtained.

Table 4. Results of Confusing Analysis

Quality of Swag	Number of Items	Number of Questions	Percentage
Very good	6, 11, 14, 20, 21, 24, 25, 27, 28	9	30%
Good	1, 2, 4, 5, 7, 12, 17, 18, 23, 30	10	33.33%
Enough	3, 8, 9, 10, 13, 16, 19, 26,	8	26.67%
Bad	22	1	3.33%
Very bad	15, 28	2	6.67%

Table 4 shows that the items most (90%) of the cheaters functioned properly. Few (10%) of cheaters are not functioning properly. Seen from the effectiveness of the distractor, this math test question is included in a good quality question compared to the research results of research by Tilaar and Hasriyanti (2019), which has 23.33% distractors in the less category and 20.00% in the poor category. Thus, most of the items in this math test have choices of answers that functioned properly.

According to Arikunto (2013), it was followed up after analyzing the items in terms of the effectiveness of the deceiver, namely: cheaters are accepted because they are categorized as good. This means that all distractors on an item have been chosen by 5% of the total number of test-takers; the bully is rewritten for not being good. This means that the deceiver has not performed its function properly (chosen by less than 5% of test-takers) and was rejected because the cheater is not good. This means that the test taker is not chosen at all (0%).

Quality of Question Items

After analyzing the items based on the validity, reliability, difficulty level, discriminating power, and tricking effectiveness, the next step is to draw conclusions on the quality of each item. In drawing these conclusions, the researchers use four aspects: validity, difficulty level, discriminating power, and deceptive effectiveness. The reliability aspect is to determine the level of consistency or consistency of an item as a whole. Therefore, reliability is not used in determining the quality of each item.

Table 5. Distribution of Overall Question Item Quality Analysis Results

The Quality of the Items	Number of Items	Number of Questions	Percentage
Good	1, 4, 5, 6,7, 9, 11, 14, 17, 18, 20, 21, 25, 27, 29	15	50%
Enough	2, 3, 8, 12, 13, 19, 23, 26, 28, 30	10	33.33%
Bad	10, 15, 16, 22, 24	5	16.67%

Based on the results of the analysis of the quality of the mathematics test questions as a whole, it was found that most (83.33%) items were classified as good quality questions by fulfilling the four criteria for good items, namely valid, had a moderate level of difficulty, had good discriminating power and trickster works fine. This is supported by the results of the reliability analysis with the reliability coefficient of the test of 0.90, which means very high.

As for the items that are not of good quality or fail, the cause should be traced. This is very useful for teachers in compiling questions in order to get good quality. Table 6 presents the causes of the failure of the mathematics test items seen from the aspects of validity, difficulty level, discriminating power, and deceptive effectiveness.

Table 6. Causes of Question Item Failure

The Cause of the Failure of the Item	Number of Items	Number of Questions	Percentage
Validity	30	1	3.33%
Level of difficulty	2, 8, 10, 12, 15, 16, 22, 23, 24	9	30%
Discernment	10, 24	2	6.67%
Deceptive Effectiveness	15, 16, 22, 28	4	13.33%

From Table 6, it can be seen the cause of the drop in good quality items. Dissimilarity is not recommended because the difficulty index is either too low or too high (Rahmat et al., 2020). This means items that are too difficult or too easy cannot discriminate between smart students and less smart students, so they do not have good discriminating power. In addition, discriminatory power can be affected by distractors. The distractor is said to be effective if the test taker chooses it from the lower group. Otherwise, if the test taker chooses it from the upper group, it means that the distractor is not functioning properly. In other words, the test items cannot discriminate between smart and less smart students.

CONCLUSION

In terms of the quality of the question items contained in the Mathematics Module of Class XI SMKN 3 Kuningan book, the results of the analysis of the samples representing each subject are most (96.67%) of the items declared valid in content by the expert. This means that most of the items are in accordance with the competency indicators to be measured. The test has very high reliability (0.90). This means that the test results meet a high level of consistency. The items have the ideal difficulty level. Most of the questions (70%) had medium

https://doi.org/10.21831/pep.v25i1.39174

difficulty, a few questions (6.67%) were very easy, and a few questions (20%) were difficult, and (3.3%) were very difficult. Most (83.33%) items had good discriminating power. Only a few questions (16.67%) had poor differentiation. This means that most of the questions were able to distinguish high and low-ability students. Most (90%) of the items had functional answer choices. Only a few questions (10%) had the answer choices not working properly. Overall, this study can be concluded that the Mathematics Test Questions at SMKN 3 Kuningan are of good quality.

REFERENCES

Arifin, Z. (2013). *Evaluasi pembelajaran*. Remaja Rosdakarya.

Arikunto, S. (2013). *Dasar-dasar evaluasi pembelajaran* (2nd ed.). Bumi Aksara.

Azwar, S. (2019). *Reliabilitas dan validitas* (4th ed.). Pustaka Pelajar.

Baskoro, E. P. (2018). *Perencanaan pelaksanaan dan evaluasi pembelajaran*. Eduvision.

Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analisis butir soal ujian Matematika kelas VII semester ganjil tahun pelajaran 2017/2018. *Mathema: Jurnal Pendidikan Matematika*, *2*(1), 57–66. https://doi.org/10.33365/jm.v2i1.459

Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF)*, *2*(1), 59–70. https://doi.org/10.24853/elif.2.1.59-70

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

Jannah, R., Hidayat, D. N., Husna, N., & Khasbani, I. (2021). An item analysis on multiple-choice questions: A case of a junior high school English try-out test in Indonesia. *Leksika: Jurnal Bahasa, Sastra Dan Pengajarannya*, *15*(1), 9–17. https://doi.org/10.30595/lks.v15i1.8768

Kusaeri, K. (2014). *Acuan dan teknik penilaian proses dan hasil belajar dalam kurikulum 2013*. Ar Ruzz Media.

Kusaeri, K., & Suprananto, S. (2012). *Pengukuran dan penilaian pendidikan*. Graha Ilmu.

Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *REiD (Research and Evaluation in Education)*, *4*(1), 70–78. https://doi.org/10.21831/reid.v4i1.20202

Maryanes, F., Fitriati, F., & Salmina, M. (2018). Analisis kualitas butir soal ujian akhir semester mata pelajaran Matematika kurikulum 2013 kelas VII SMP Negeri 8 Banda Aceh. *Prosiding Seminar Nasional Pendidikan Dasar 2018*, 576–580. https://repository.bbg.ac.id/handle/737

Rahayu, R., & Djazari, M. (2016). Analisis kualitas soal pra ujian nasional mata pelajaran Ekonomi Akuntansi. *Jurnal Pendidikan Akuntansi Indonesia*, *14*(1), 84–94. https://doi.org/10.21831/jpai.v14i1.11370

Rahmat, I., Masi, L., & Anggo, M. (2020). Analisis butir soal Ujian Akhir Semester (UAS) di masa pandemi pada mata pelajaran Matematika untuk tahun ajaran 2019/2020 kelas VII SMPIT Bina Insan Mandiri Al-Masrur Kendari. *Jurnal Penelitian Pendidikan Matematika*, *8*(3), 491–504. https://doi.org/10.36709/jpp.v8i3.16624

Ratnawulan, E., & Rusdiana, A. (2015). *Evaluasi pembelajaran*. Pustaka Setia.

https://doi.org/10.21831/pep.v25i1.39174

Setiawan, D. F. (2018). *Prosedur evaluasi dalam pembelajaran*. Deepublish.

Suharman, S. (2018). Tes sebagai alat ukur prestasi akademik. , . Retrieved from. *At-Ta'dib: Jurnal Ilmiah Prodi Pendidikan Agama Islam*, *10*(1), 93–115. https://ejournal.staindirundeng.ac.id/index.php/tadib/article/view/138

Sukanti, S. (2011). Penilaian afektif dalam pembelajaran Akuntansi. *Jurnal Pendidikan Akuntansi Indonesia*, *9*(1), 74–82. https://doi.org/10.21831/jpai.v9i1.960

Sumiati, A., Widiastuti, U., & Suhud, U. (2018). Workshop teknik menganalisis butir soal dalam meningkatkan kompetensi guru di SMK Cileungsi Bogor. *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *2*(1), 136–153. https://doi.org/10.21009/JPMM.002.1.10

Supriadi, W. O. S., Rahim, U., & Zamsir, Z. (2019). Kualitas tes sumatif mata pelajaran Matematika kelas VIII semester genap SMP Negeri 20 Kendari tahun pembelajaran 2016/2017. *Jurnal Penelitian Pendidikan Matematika*, *6*(3), 85. https://doi.org/10.36709/jppm.v6i3.9142

Tilaar, A. L. F., & Hasriyanti, H. (2019). Analisis butir soal semester ganjil mata pelajaran Matematika pada sekolah menengah pertama. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, *8*(1), 57–68. https://doi.org/10.15408/jp3i.v8i1.13068

Wijayanti, P. S. (2020). Item quality analysis for measuring mathematical problem-solving skills. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *9*(4), 1223–1234. https://doi.org/10.24127/ajpm.v9i4.3036

Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://doi.org/10.18592/tarbiyah.v7i1.2100

Zein, M., & Darto, D. (2012). *Evaluasi pembelajaran matematika*. Daulat Riau.

Zuhera, Y., Habibah, S., & Mislinawati, M. (2017). Kendala guru dalam memberikan penilaian terhadap sikap siswa dalam proses pembelajaran berdasarkan kurikulum 2013 di SD Negeri 14 Banda Aceh. *Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *2*(1), 73–87. http://www.jim.unsyiah.ac.id/pgsd/article/view/2534