Jurnal Math Educator Nusantara
Wahana publikasi karya tulis ilmiah di bidang pendidikan matematika
p-issn: 2459-9735 e-issn: 2580-9210
http://ojs.unpkediri.ac.id/index.php/matematika

Item analysis of mathematical literacy test and self efficacy of learners

Eka Mega Nanda[1*], Kismiantini[2] , Elly Arliani[3]
[1]Mathematics Education Study Program, Yogyakarta State University. Jalan Colombo, 55281, Yogyakarta, Indonesia.
E-mail:[1] ekamega.2023@student.uny.ac.id, [2] kismi@uny.ac.id ,[3]Arlianielly@uny.ac.id

Article received : October 11, 2025.
Article revised : November 24, 2025.
Article Accepted: November 25, 2025.
** Corresponding author*

Abstract: *This study aims to analyze the quality of daily test items and non-test instruments (questionnaires) on Two-Variable Linear Equation Systems in junior high school, in terms of validity, reliability, discrimination power, and difficulty level. The method used was descriptive analysis with 64 eighth-grade students (academic year 2024/2025) as subjects. The findings show that all instruments are valid (essays are very high, multiple-choice questions and questionnaires are high), but the reliability of exam questions (essays and multiple-choice questions) is still in the moderate category. Further analysis of multiple-choice questions shows a predominance of easy difficulty levels (92%) and a number of items (25%) with poor discrimination power, indicating the need for item revision.*
Keywords: Validity; Reliability; Distinguishing Power; Level of Difficulty.

Analisis butir soal tes literasi matematika dan efikasi diri peserta didik

Abstrak: Penelitian ini bertujuan untuk menganalisis kualitas item tes ujian hari dan dan instrumen non-ujian (kuesioner) pada materi Sistem Persamaan Linear Dua Variabel di SMP, ditinjau dari aspek validitas, reliabilitas, daya diskriminasi, dan tingkat kesukaran. Metode yang digunakan adalah analisis deskriptif dengan subjek 64 peserta didik kelas VIII (T.A 2024/2025). Temuan menunjukkan bahwa semua instrumen valid (esai tingkat tinggi, pilihan ganda dan kuesioner tinggi), namun reliabilitas soal ujian (esai dan pilihan ganda) masih berada dikategori sedang. Analisis lebih lanjut pada soal pilihan ganda menunjukkan dominasi tingkat kesulitan yang mudah (92%) dan adanya sejumlah item (25%) dengan daya deskriminasi yang tidak baik, mengindikasikan perlunya revisi soal.
Kata Kunci: Validitas; Reliabilitas; Daya Beda; Tingkat Kesukaran.

INTRODUCTION

Learning evaluation is a structured, continuous and comprehensive activity to control and determine the quality (value and meaning) of learning related to various learning elements by considering and using the performance set as an accountability guide in carrying out the learning process (Ropii & Fahrurrozi, 2017). Learning evaluation covers all aspects of learning, while assessment is the process of determining the value or quantity of an object. Retnoningsih (2020) defines assessment as an organized process that gathers, examines, and interprets data to determine whether or not student have met learning objectives. The assessment's goal is to gather information that will be utilized to enhance both the learning process and the learning that has already been completed. Assessment covers various aspects such as cognitive, affective and psychomotor, as well as with various techniques such as tests, non tests and portfolios (Badrudin et al., 2024). The assessment that will be carried out is applying test techniques and non-test techniques to student learning outcomes.

CITATION FORMATS: Nanda, E. M., Kismiantini, K., & Arliani, E. (2025). Item analysis of mathematical literacy test and self efficacy of learners. *Jurnal Math Educator Nusantara: Wahana Publikasi Karya Tulis Ilmiah Di Bidang Pendidikan Matematika*, *11*(2), 307-318. https://doi.org/10.29407/jmen.v11i2.27281

Test technique is an evaluation method using measuring instruments that have objective standards, allowing measurement and comparison of psychological conditions or individual behavior (Matondang, 2009). Arifin (2012) states that a test is a tool consisting of questions that must be done by students to measure a certain aspect of behavior, so that the test becomes a measuring instrument. According to Arikunto (2009), If the test satisfies the standards, it can be considered a good measurement tool, namely: validity, reliability, objectivity, and practicability. One form of assessment of the test technique that will be applied in this study is the assessment of the daily mathematics test of VIII grade students on the material of the system of linear equations of two variables using multiple choice questions and essay questions.

Non-test techniques are alternative assessment methods that can be used to describe various aspects of learning and student development. Authentic assessment with non-test techniques in elementary schools can measure the honesty, responsibility, and discipline of students at home and at school (Bisri & Ichsan, 2015). To optimize overall educational outcomes and provide a comprehensive assessment of learner competencies across cognitive, emotional, and psychomotor domains, non-test evaluation methodologies must be developed (Anshari et al., 2024). One of the non-test assessments that will be applied in this study is a learner self-efficacy questionnaire.

One of the processes of going over test questions to find ones that can be used is item analysis (Nasir, 2015). Item analysis must accommodate the entire curriculum and ensure that basic competencies and competency standards are met. It is expected that item analysis will provide results regarding whether the questions given are functional or not (Nuswowati et al., 2010). According to Immanuel et al. (2024) with item analysis, teachers can find out the pattern of students' answers, identify the difficulties faced by students and assess which concepts have been understood by students. So that item analysis is an important aspect of learning. In addition to determining whether or not students have grasped to topic. Teachers can also determine the challenges that students face during the learning process, which aids in establishing a conductive learning environment and achieving positive outcomes.

A test can be considered good for several reasons. First, the validity of the test which is determined by how well the test questions can measure certain abilities. The Second is the test's dependability, which demonstrates its accuracy. The third factor is the question's difficulty level, which is determined by dividing the number of test takers who properly answered by the total number of participants. The last factor is the differential power, which is the question's capacity to identity whether or not student have understood the topic (Nur & Palobo, 2018).

Earlier research has consistently identified serious issues regarding the quality of mathematics evaluation instruments. The urgency for this study stems from the findings of Hamimi et al. (2020), which revealed that most exam questions were declared invalid because their item validity levels were still classified as very low and low. Furthermore, the test questions were considered unusable as they fell into the category of low reliability.

This fundamental problem with the quality of measurement tools is reinforced by the findings of (Tilaar & Hasriyanti, 2019), which quantitatively demonstrated the low quality of the evaluation instruments used. In their study, only 5 multiple-choice questions were categorized as very good, while 16 questions needed to be revised and 9 questions had to be discarded, indicating a systemic failure in item construction. Further specific findings supporting the need for extensive revision were presented by Halik et al. (2019), Their item analysis results indicated critical issues, particularly with the questions' differentiating power, where three things were categorized as really dreadful and twelve as bad. Additionally, the effectiveness of the alternatives was poor for three of the items. Collectively, these findings underscore the high urgency for conducting a comprehensive evaluation and fundamental improvement of the assessment instruments currently in use.

Based on several previous studies related to the item analysis of the final semester mathematics exam that has been made, it is not fully in the good category. Previous research discusses the analysis of school final exam items and daily test questions. While the research that will be conducted by researchers is the analysis of daily test items and self efficacy questionnaires for junior high school students in grade VIII.

Unlike previous studies, which mostly focused on large-scale summative instruments such as final exam questions, this study specifically evaluated the quality of daily test questions, which are routine formative evaluation instruments, and analyzed self efficacy questionnaire instruments. The quality of the evaluated based on four main criteria: validity, reliability, discriminating power and item difficulty level.

METHODS

With data sources of mathematics daily exam questions on the content of the System of Linear Equations of Two Variables (SPLDV) class VIII and the student answer keys, this study employs data gathering approaches in the form of documentation. To ascertain the problem's validity, reliability, distinction, and degree of complexity, item analysis is done. 64 junior high school students in grade VIII during the 2024/2025 academic year participated in this study. The sampling technique used in this study was purposive sampling, which is a sampling method where individuals are selected based on certain predetermined criteria (Cochran & Wiley, 1977). The purpose of purposive sampling is to ensure that the selected group meets certain criteria relevant to the study.

The questions for the math daily test assessment in class VIII are questions made by the author. The problem consists of two types of questions, namely multiple choice and essay. Multiple choice questions consisting of 10 items with four answer choices and essay questions consisting of 3 items and a questionnaire of students' self-efficacy. This research begins with collecting questions, answer keys, and answers from students. After that, give a score or assessment of the answer to the question. Furthermore, the researcher entered the students' answers and answer keys into the Anbuso 8.0 program after the results came out, the researcher analyzed the data so that he would get the results of the differential power and the level of difficulty of the items.

Data Analysis

Data analytical techniques are used to assess the questions' validity, reability, uniqueness and level of difficulty. The validity analysis in this study used Microsoft excel, while the reliability analysis used R studio software and the differentiation of questions and the difficulty level of questions using Anbuso 8.0 software.

Validity Analysis

All types of questions, namely multiple choice, essay and student questionnaire were tested for validity using Microsoft excel. Decision criteria according to Sutama (2014) as follows.

Table 1. Criteria for Validity Coefficient

Validity Criteria	Description
$0{,}79 < VC \leq 1{,}00$	Very High
$0{,}59 < VC \leq 0{,}79$	High
$0{,}39 < VC \leq 0{,}59$	Medium
$0{,}19 < VC \leq 0{,}39$	Low
$0{,}00 < VC \leq 0{,}19$	Very Low

Reliability Analysis

The word "reliability" refers to the degree of dependability of measurement results. A measurement result can be regarded reliable if it is conducted on the same group of subjects, gave results that are essentially the same, and the features of the issue that were measured have not changed Matondang (2009).

Table 2. Reliability for Validity Coefficient

Reliability Coefficient	Description
$r \leq 0{,}20$	Very Low
$0{,}20 < r \leq 0{,}40$	Low
$0{,}40 < r \leq 0{,}60$	Medium
$0{,}60 < r \leq 0{,}80$	High
$0{,}80 < r \leq 1{,}00$	Very High

Distinguishing Power

Differentiating power is the capacity to distinguish between students with high and low ability. The function of the distinguishing power is to find out individual differences in detail (Iskandar & Rizal, 2018). If an item's index of distinguishing power is low, it will not be able to differentiate between students with high and low ability levels. When using Content-Referenced Measures in test analysis, the item discriminating power index is not considered to be very significant as long as it is not negative (Ebell & Friesbie, 1991).

Table 3. Criteria for Distinguishing Power

Criteria	Description
0,40 –1,00	Good
0,30 – 0,39	Medium
0,20 – 0,29	Good Enough
Negatif – 0,19	Poor

Source:Iskandar & Rizal (2018)

Level of Difficulty

The level of difficulty of the question is a numerical description of the difficulty level of a question (Fietri et al., 2021). Correcting students' answers sheets is one way to determine the items in multiple choice questions. Whereas the incorrect response receives a value of 0, the right response receives a value of 1. The difficulty index number is calculated using formula (1):

$$P_i = \frac{\Sigma X_i}{N} \qquad (1)$$

Description:

P_i : the level of difficulty of the question item

ΣX_i : number of students answering correctly

N : number of all students

The table 4 displays the requirements for the degree of difficulty.

Table 4. Criteria for Level of Difficulty

Hardness Criteria	Category
p < 0,30	Difficult
0,30 $\leq p \leq$ 0,70	Moderate
p > 0,70	Easy

Source: Iskandar & Rizal (2018)

Good questions are those that are neither too easy nor too difficult. If the question is too difficult, students will be lazy to do it and not motivated to do it again because it is impossible to achieve (Ambarwati & Ismiyati, 2022).

RESULT AND DISCUSSION

Research Result

Results and discussion in this study by collecting question data, questionnaires and student answers. Each data is reviewed and processed using the help of excel, R Studio and Anbuso, so that the results obtained will be detailed below.

Validity Analysis

Validity Test of Multiple Choice Questions

Table 5. Test Results Decision Criteria

Aiken Index	0,646
Description	High Validity

Table 5 shows that in order to earn a high validity category, the Aiken index of the validity test of multiple choice questions received a value of 0.646.

Validity Test of Essay Questions

Table 6. Result of the Essay Question Validity Test

Aiken Indek	0,917
Description	Very High Validity

According to table 6, the essay question can be classified as having very high validity because the Aiken index achieved is 0.917.

Validity Test of Student *Self Efficacy* Questionnaire

Table 7. Questionnaire Validity Test Results

Aiken Indek	0,656
Description	High Validity

Table 7 indicates that the students' self efficacy questionnaire yielded an Aiken index of 0.656, indicating that the questionnaire's validity falls into the high range.

Reliability Analysis

Table 8 displays the reliability test result for the multiple choice questions with the help of Excel.

Table 8. Multiple Choice Problem Reliability Test Results

Number of Variants	3,615079
Total Variant	1,541667
Decision	0,625686

Table 8 shows that the decision criterion is 0.62, indicating that it is classified as fairly reliable. R studio software was used to acquire the findings of the essay question reliability test and the srudents' self efficacy questionnaire, which are described in full below.

Table 9. Results of Essay Problem Reliability Test

Test	raw_alpha
Cronbach alpha	0,65

Table 9 shows that the reliability test for the essay question yielded a result of 0.65, indicating that the value falls into the high reliability group.

Table 10. Reliability Test Results of Self Efficacy Questionnaire

Test	raw_alpha
Cronbach alpha	0,7

Table 10 shows that the reliability test of the students' self efficacy questionnaire obtained a result of 0.7, which means that the value is in the high reliability category.

Mathematical Literacy Test Analysis

Distinguishing Power

The following differentiated power findings are obtained from the data analysis of the math daily test scores for the eighth grade on the content of the system of linear equations of two variables.

Table 11. Results of Differentiability Analysis of Multiple Choice Questions

No Item	Distinguishing Power	
	Coefficient	Description
(1)	(2)	(3)
1	0,596	Good
2	0,313	Good
3	0,396	Good
4	0,460	Good
5	0,286	Good Enough
6	0,117	Not Good
7	0,000	Not Good
8	0,218	Good Enough
9	0,313	Good
10	0,069	Not Good
11	0,374	Good
12	0,220	Good Enough

Table 11 shows that accepted questions are good, with a total of 6 items (50%), quite good questions totaling 3 items (25%), and bad questions totaling 3 items (25%). The findings of the analysis of the class VIII daily test questions' distinguishing power for the 2024/2025 academic year are shown in Figure 1.

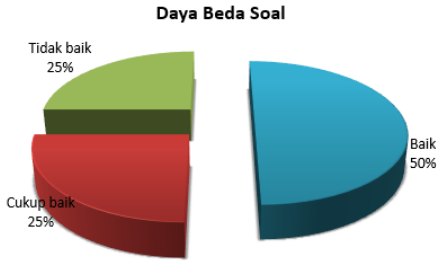

Figure 1. Distinguishing Power of Multiple Choice Questions

Table 12. Results of Differentiability Analysis of Essay Questions

Item	Distinguishing Power	
No.	Coefficient	Description
(1)	(2)	(3)
1	0,459	Good
2	0,411	Good
3	0,507	Good

Based on table 12 shows that the three questions are accepted as good (100%). The diagram of the results of the analysis of the differentiability of the essay questions of the 8th grade daily test on the material of the System of Linear Equations of Two Variables (SPLDV) in the 2024/2025 school year can be seen in Figure 2.

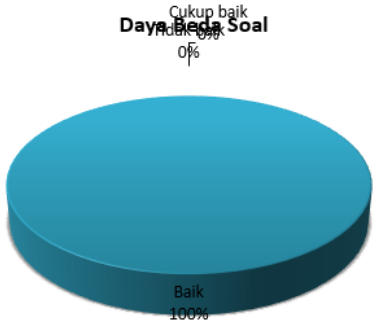

Figure 2. Differentiability of Essay Questions

Diffuculty Level

Table 13. Results of Analysis of the Level of Difficulty of Multiple Choice Questions

Item No.	Level of Difficulty	
	Coefficient	Description
(1)	(4)	(5)
1	0,938	Easy
2	0,813	Easy
3	0,875	Easy
4	0,781	Easy
5	0,938	Easy
6	0,844	Easy
7	1,000	Easy
8	0,469	Medium
9	0,813	Easy
10	0,844	Easy
11	0,844	Easy
12	0,781	Easy

Table 13 indicates that 11 items (92%) fall into the easy group, while 1 item (8%), falls into the medium category. Figure 3 shows the findings of the analysis of the multiple choice question difficulty level on the class VIII daily test of the System of Liniear Equations of Two Variables (SPLDV) material for the 2024/2025 academic year.

Figure 3. Level of Difficulty of Multiple Choice Questions

Table 14. Results of Analysis of the Level of Difficulty of Essay Questions

Item No	Level of Difficulty	
	Coefficient	Description
(1)	(4)	(5)
1	0,547	Medium
2	0,443	Medium
3	0,479	Medium

Table 14 shows that each item's essay questions fall into the medium group in terms of difficulty. Figure 4 displays the findings of the analysis of the essay questions' degree of difficulty on the class VIII daily assessments covering the System of Linear Equations of Two Variables (SPLDV) for the 2024/2025 academic year.

Figure 3. Level of Difficulty of Essay Questions

Discussion

According to the aforementioned data analysis, the author's question items were still in a low state. This is consistent with the findings of Lestari et al. (2023), the analysis of the discriminating power has a index that states that it is lacking, so that the three questions are not suitable for use and must be revised so that thet can be used.

Research conducted by Balau et al. (2021), Test validity indicates that most questions do not meet the expected validity standards. The discriminatory power shows that there are 22 questions in the poor category, meaning they cannot distinguish between students' abilities. There are still a lot of question in the easy and medium difficulty categories and none of them are challenging. The test questions created by teachers at Matabulu State Junior High School are still not in good shape and need to be revised.

The student self efficacy questionnaire receives a score of 0.65 with a high validity category, essay questions receive a score of 0.91 with a very high validity category, and multiple choice questions receive a score of 0.64 with a high validity category. Multiple choice questions receive a medium category reliability score 0f 0.62, essay questions receive a medium category reliability score of 0.65, and the questionnaire's reliability receives a high category score of 0.7. These positive findings stand in stark contrast to previous research that has highlighted pervasive quality issues. For instance Hamimi et al. (2020) reported that most prior exam questions were declared invalid due to very low item validity and low reliability, rendering them unusable.

The items at numbers 1, 2, 3, 4, 9 and 11 have strong distinguishing power, whereas the items at numbers 5, 8, and 12 have pretty good differentiating power. This means that the questions in the good and good enough category on the differentiating power of the questions have been able to distinguish the level of differentiating power of students whose knowledge is high and low. In question items number 6, 7, and 10 are categorized as not good so that these items should not be used again for testing or further tests. According to Iskandar & Rizal (2018) in their research found that in terms of differentiating power there are 4 items (5%) that have very good differentiating power which means that the item can distinguish between students who are good and less good, 20% who have differentiating power in the good category, 11 items (13.75%) have good enough differentiating power which means they have to go through the improvement stage so that they can be reused, and 49 items (61.25%) have poor differentiating power so they must be discarded.

This Positive finding directly contrasts with research such a Halik et al. (2019), who found as substantial number of item (3 really dreadful and 12 bad) in their test to possess low differentiating power. Althought item 6, 7, and 10 were categorized as 'not good' and should be revised or excluded, the overall hight performance of the instrument indicates that the majority of the developed items are effective and functional.

Because the multiple choice choices were so complicated, there was only question in the medium category and eleven in the easy category. As for the essay items, it was found that 3 essay questions were in the medium category. Hanna & Retnawati (2022) revealed that good test questions should not have questions with the lowest or highest difficulty levels. A good question is one that is neither too simple noor too complex. Therefore, analyzing the quality of the items needs to be done to determine the feasibility of items to be used in tests or in research.

CONCLUSIONS

Based on the theoretical investigation and data analysis findings, it can be said that the student self efficacy questionnaire has a high validity category, essay questions have a very high validity category, and multiple choice questions have a high validity category. The questionnaire's high level dependability, essay questions' medium level reliability, and multiple choice questins' medium level reliability. Differentiation by multiple choice three things are classified as fairly good, three are classified as not good, and six are classified as good. One question's answers fall into the medium group for the multiple choice items' level of difficulty, while the remaining eleven fall into the easy category.

Some research recommendations based on the completed study review are as follows: 1) When creating exam questions, it is important to ensure that they are legitimate, reliable, differentiated, and challenging before distributing them to students. 2) Future researchers can examine the validity of multiple-choice questions. Items with low discriminating power should be revised immediately by immproving the function of distractors and reformulating the test composition to increase the proportion of medium and difficult questions in order to achieve better discrimination. 3) Finally, further research is also recommended to empirically explore the

relationship between students self efficacy and their performance on various types of evaluation formats, to understand the interaction between psycological factors and academic outcomes.

REFERENCES

Ambarwati, Y. F., & Ismiyati, I. (2022). Analisis Butir Soal Pilihan Ganda Ulangan Akhir Semester Genap Mata Pelajaran Kearsipan. *Measurement In Educational Research (Meter*, *1*(2). https://doi.org/10.33292/meter.v1i2.144

Anshari, A., Hibatullah, M. Z., & Widyanti, E. (2024). Pengembangan Evaluasi Teknik Non Tes. *Guruku: Jurnal Pendidikan Dan Sosial Humaniora*, *2*(3), 149–161. https://doi.org/10.59061/guruku.v2i3.702

Arifin, Z. (2012). *Evaluasi Pembelajaran (Edisi revisi)*. Direktorat Jenderal Pendidikan Islam, Kementerian Agama RI.

Arikunto, S. (2009). *Dasar-Dasar Evaluasi Pendidikan (Edisi Revi*. PT Bumi Aksara.

Badrudin, R., R, M., Rahmi, R. S., & Mulyani, S. (2024). Pengembangan Manajemen Penilaian Pendidikan. In *JIIP (Jurnal Ilmiah Ilmu Pendidikan* (Vol. 7).

Balau, M., Pesik, A., & Damai, I. W. (2021). Analisis Kualitas Butir Soal Buatan Guru Mata Pelajaran Matematika Kelas VIII SMP Negeri Satap Matabulu Kabupaten Bolaang Mongondow Timur. *MARISEKOLA: Jurnal Matematika Riset Edukasi Dan Kolaborasi*, *2*(1), 13–18.

Bisri, H., & Ichsan, M. (2015). Penilaian Otentik dengan Teknik Nontes di Sekolah Dasar. *Jurnal Sosial Humaniora*, *6*(5), 81–93.

Cochran, W. G., & Wiley, J. (1977). *Sampling Techniques (Third edit*. John Wiley & Sons, Inc.

Ebell, R. L., & Friesbie, D. A. (1991). Essentials of Educational Measurement. *Journal of Educational Measurement*. https://doi.org/https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a-_frisbie_essentials_of_edbookfi-org.pdf

Fietri, W. A., Zulyusri, & Violita. (2021). Analisis Butir Soal Biologi Kelas XI Madrasah Aliyah Skinah Kerinci Menggunakan Program Komputer Anates 4.0 For Windows. *Natural Science: Jurnal Penelitian Bidang IPA Dan Pendidikan IPA*, *7*(2). https://doi.org/10.35791/agrsosek.17.2.2021.33834

Halik, A. S., Mania, S., & Nur, F. (2019). Analisis Butir Soal Ujian Akhir Sekolah (Uas) Mata Pelajaran Matematika Pada Tahun Ajaran 2015/2016 Smp Negeri 36 Makassar. *Al Asma : Journal of Islamic Education*, *1*(1). https://doi.org/10.24252/asma.v1i1.11249

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). https://doi.org/10.33365/jm.v2i1.459

Hanna, W. F., & Retnawati, H. (2022). Analisis Kualitas Butir Soal Matematika Menggunakan Model RASCH Dengan Bantuan Software QUEST. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *11*(4), 3695–3704. https://doi.org/10.24127/ajpm.v11i4.5908

Immanuel, C., Manik, S. D. P., Nababan, A. P., Sianturi, B. Y., & Hasnah, A. (2024). Analisis Butir Soal Pilihan Ganda Mata Pelajaran Matematika Materi Bangun Datar Berbasis Budaya LOkal Di SDN 106163 Bandar Klippa. *Jurnal Ilmiah Multidisiplin Terpadu*, *8*(6), 323–331.

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

Lestari, A. S., Fitrianna, A. Y., & Zanthy, L. S. (2023). Analisis Butir Soal Tes Materi Sistem Persamaan Linear Dua Variabel Pada Siswa Kelas VIII. *Pembelajaran Matematika Inovatif*, *6*(1), 367–376. https://doi.org/10.22460/jpmi.v6i1.12389

Matondang, Z. (2009). Validitas dan Reliabilitas Suatu Instrumen Penelitian. *JURNAL TABULARASA PPS UNIMED*, *6*(1), 87–97.

Nasir, M. (2015). Analisis Empirik Program Analisis Butir Soal Dalam Rangka Menghasilkan Soal Yang Baik dan Bermutu Sebagai Alat Evaluasi Pembelajaran Fisika. In *Prosiding Semirata* (pp. 336–347).

Nur, A. S., & Palobo, M. (2018). Pelatihan Analisis Butir Soal Berbasis Komputerisasi Pada Guru SD. *MATAPPA: Jurnal Pengabdian Kepada Masyarakat*, *1*(1), 5–11. https://www.academia.edu/download/92970935/45

Nuswowati, M., Binadja, A., Efti, K., & Ifada, N. (2010). Pengaruh Validitas Dan Reliabilitas Butir Soal Ulangan Akhir Semester Bidang Studi Kimia Terhadap Pencapaian Kompetensi. *Jurnal Inovasi Pendidikan Kimia*, *4*(1), 566–573.

Retnoningsih, E. (2020). Model-Model Dan Alat Dalam Penilaian Di PKN SD dan MI. *Osf*, *1*(1). https://doi.org/10.31227/osf.io/gn9df

Ropii, M., & Fahrurrozi, M. (2017). Evaluasi Hasil Belajar.

Sutama, I. M. (2014). *Statistika Pendidikan*. UNS Press.

Tilaar, A. L., & Hasriyanti. (2019). Analisis Butir Soal Semester Ganjil Mata Pelajaran Matematika Pada Sekolah Menengah Pertama. *8*, 57–68. https://doi.org/10.15408/jp3i.v8i1.13068