Beta: Jurnal Tadris Matematika, 19 (1) 2026: 39-57
DOI 10.20414/betajtm.v19i1.779

Research articles

Assessing the problem-solving test items and self-efficacy questionnaire in statistics education

Imam Nursahid[1], Agus Maman Abadi[1], Wahyu Anugerah Sianipar[1]

Abstrak Penelitian ini bertujuan untuk menganalisis kualitas butir soal kemampuan pemecahan masalah dan angket *self-efficacy* dengan memanfaatkan aplikasi AnBuso sebagai alat bantu analisis. Soal kemampuan pemecahan masalah dikembangkan untuk mengukur kemampuan kognitif siswa pada level memahami (C2) hingga mengevaluasi (C5) dalam konteks materi statistika. Metode penelitian ini menggunakan pendekatan deskriptif kuantitatif dengan fokus utama pada siswa kelas VIII di SMP Negeri 1 Mlati pada materi statistika. Sampel penelitian terdiri atas 32 orang siswa yang dipilih sebagai subjek penelitian. Proses analisis dilakukan dari segi karakteristik soal yaitu daya pembeda, tingkat kesulitan kemudian validitas ahli, dan reliabilitas soal. Temuan penelitian menunjukkan bahwa instrumen pemecahan masalah memiliki daya pembeda yaitu 53.33% soal memiliki fungsi pembeda yang baik, 20% cukup baik, dan 26.67% tidak baik. Sebanyak 33.33% soal tergolong mudah, 66.67% sedang, dan tidak ada soal yang tergolong sulit. Kevalidan sebesar 0.90 (sangat valid) dan reliabilitas soal pilihan ganda sebesar 0.51 (sedang), soal uraian 0.72 (tinggi). Sementara angket *self-efficacy* memiliki kevalidan sebesar 0.89 (sangat valid) dan reliabilitas sebesar 0.83 (sangat tinggi). Secara keseluruhan, instrumen tes dan angket yang digunakan menunjukkan tingkat kevalidan yang sangat valid serta reliabilitas yang memadai. Hasil analisis ini diharapkan dapat menjadi acuan bagi pendidik dan pengembang kurikulum dalam menyusun alat evaluasi yang lebih baik, sehingga tujuan pembelajaran dapat tercapai secara optimal.

Kata kunci *Analisis butir, Kemampuan pemecahan masalah, Self-efficacy, Statistika*

Abstract The study aims to analyze the quality of problem-solving test items and a self-efficacy questionnaire using the AnBuso application as an analytical tool. The problem-solving test items were developed to measure the students' cognitive abilities ranging from understanding (C2) to evaluating (C5) in statistical topics. This study employed a quantitative descriptive approach focusing on eight-grade students at a public school SMP Negeri 1 Mlati. The sample consisted of 32 students selected as research subjects. The analysis examined the test item characteristics, including discriminative power, difficulty level, expert validity, and item reliability. The findings indicated that 53.33% of the items had good discriminative function, 20% had fairly good function, and 26.67% had poor function. A total of 33.33% of the items were classified as easy, 66.67% as moderate, and none as difficult. The validity was 0.90 (highly valid), and the reliability of multiple-choice items was 0.51 (moderate), while open-ended items reached 0.72 (high). Meanwhile, the self-efficacy questionnaire had a validity of 0.89 (highly valid) and a reliability of 0.83 (very high). Overall, the test instruments and questionnaire showed high validity and adequate reliability. It is hoped that the research finding can serve as a reference for educators and curriculum developers in designing better evaluation tools so that the learning objectives can be optimally achieved.

Keywords *Item analysis, Problem-solving ability, Self-efficacy, Statistics*

[1]**Universitas Negeri Yogyakarta, Yogyakarta, Indonesia,** imamnursahid.2024@student.uny.ac.id

Introduction

Authentic assessment in education involves the comprehensive measurement of learning outcomes in the affective, cognitive, and psychomotor domains as a representation of student competencies. The cognitive domain in the context of mathematics learning is a primary focus because it is directly related to the thinking processes involved in understanding concepts, applying procedures, analyzing situations, and evaluating solutions (NCTM, 2000). Based on the revised Bloom's taxonomy developed by Anderson and Krathwohl (2001), the cognitive domain consists of six levels of thinking processes: remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6). This study focuses on measuring mathematical problem-solving abilities at the levels of understanding (C2) through evaluating (C5) because problem-solving activities require conceptual understanding, the application of strategies, the analysis of relationships among information, and the evaluation of the accuracy of solutions. The success of an educational objective can be measured through the implementation of evaluations designed systematically and in alignment with the characteristics of the cognitive abilities being measured. These evaluations are part of the measurement and assessment processes conducted on students. An assessment cannot be conducted before a measurement.

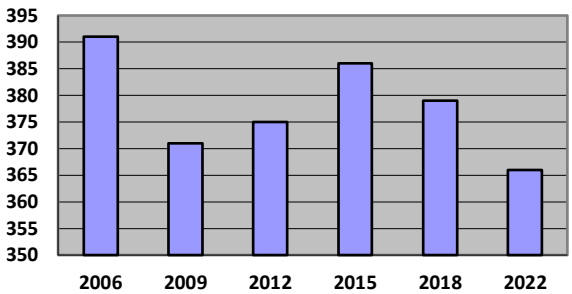
Clearly, evaluation is defined as a process of collecting data and based on that data, determining the extent to which educational objectives have been implemented or achieved. If they have not been achieved, the evaluation must explain the causes and outline the next steps to be taken. The results are presented in the form of a structured report, supplemented with visual graphs, and are freely available (Arikunto, 2019). Evaluation serves as feedback aimed at improving education (Sudijono, 2017). Educational evaluation and psychometrics, from a scientific perspective, do not merely involve data collection but constitute a systematic process encompassing measurement, assessment, and decision-making based on empirically analyzed data through classical test theory or item response theory approaches (Mardapi, 2012; Istiyono, 2020; Retnawati, 2016). Therefore, the quality of evaluation results is largely determined by the quality of the instruments used. A good instrument must meet the characteristics of validity and reliability so that the measurement results can be trusted and used as a basis for educational decision-making (Kartowagiran, 2012).

Based on this evaluation, it can be measured on the extent to which learning objectives have been achieved and identify both the strengths and weaknesses of the learning process. This allows us to address any shortcomings in the learning process while maintaining the aspects that students find engaging. Additionally, to assess quality, the test items that have been designed need to be analyzed. This is done to determine whether each test item is suitable for use or not through item analysis (Okyranida, Mayanty, & Widiyatun, 2024). Furthermore, by testing the items, the quality of those items can be assessed to the extent that they measure various aspects of the test battery being administered (Arifin, 2012). Based on these findings, this item analysis is capable of measuring students' abilities accurately and precisely, resulting in a valid instrument for use that aligns with modern psychometric principles (Retnawati et al., 2017).

In the 21st century, the world faces increasingly complex challenges, including issues of survival and challenges in the education sector. This century requires every individual to master the essential skills needed to achieve success in life. Education is expected to prepare students to master these various skills. These 21st-century skills align with the four pillars of education outlined by UNESCO: knowing, doing, living together, and growing. In the context of education,

there are four main 21st-century skills known as the 4Cs: critical thinking and problem solving, creative thinking, communication, and collaboration (Astuti, Aziz, Sumarti, & Bharati, 2019).

One of the keys 21st-century skills is the ability to solve problems. This ability is crucial in mathematics and needs to be practiced and developed by students in this era (Florea & Hurjui, 2015). Problem-solving involves steps in which students apply the knowledge, rules, techniques, skills, and concepts they have previously learned to find solutions. Problem-solving ability is a clear indication of the benefits and relevance of mathematics. Indicators of problem-solving ability include understanding mathematical problems, planning mathematical solutions, solving problems according to a plan, and reviewing solutions (Syaiful, Kamid, Muslim, & Huda, 2020). Problem-solving helps students expand their understanding of mathematics and develop skills to tackle challenges in daily life (Căprioară, 2015). Learning mathematics through problem-solving contributes to a deeper conceptual understanding (Inoue, Asada, Maeda, & Nakamura, 2019). In the context of mathematics education, problem-solving skills are a crucial element and an integral part of the curriculum in various countries (Wong & Yip, 2023). This is due to the significant role of mathematics in helping students solve problems involving numbers in their lives (NCTM, 2000). Global empirical findings increasingly underscore the urgency of strengthening problem-solving skills in the context of mathematics. Data from the Programme for International Student Assessment (PISA) indicate that Indonesian students' mathematics scores remain somewhat inconsistent.



Graph 1. Trends of the Indonesian math scores in PISA from 2006 to 2022

As displayed in Graph 1, the Indonesian's mathematics scores have not shown a significant upward trend. This situation indicates that the majority of Indonesian students remain at a proficiency level that requires strengthening in reasoning, modeling, and contextual problem-solving areas. At the junior high school level, which is the focus of this study, these findings serve as a critical signal that the learning and assessment systems must be directed toward strengthening mathematical competencies, particularly problem-solving skills. Thus, strengthening problem-solving skills in mathematics education at the junior high school level is no longer merely a curricular requirement but a strategic, evidence-based necessity for sustainably improving students' mathematics proficiency.

Problem-solving skills are one of the primary goals of mathematics education that students must achieve. Every day, various problems require problem-solving skills, whether consciously or unconsciously (Khotimah, Khoirunnisa, & Bilda, 2020). Problem-solving is one of the most important cognitive aspects used in daily life and is a crucial component of mathematics. Students must master problem-solving skills to be more thorough in solving mathematical problems related to daily life, making this skill crucial in mathematics education (Balqis, Patmawati, & Yulianto, 2024).

In addition to the cognitive aspect, another aspect that plays a role in mathematics learning is the affective aspect. Self-efficacy is one of the key factors in determining an individual's mathematics achievement, particularly in performing tasks involving problem-solving questions, and it is evident that there is a positive and mutually reinforcing relationship between problem-solving ability and self-efficacy. If a student possesses strong mathematical problem-solving skills, that student also demonstrates high self-efficacy (Nuutila, Tapola, Tuominen, Molnár, & Niemivirta, 2021). Currently, most students still fall into the low self-efficacy category. This is evidenced by a tendency to give up when encountering difficulties in learning or solving problems. To this day, many students still perceive mathematics as a difficult subject and tend to remain silent and hesitate to ask questions (Alifia & Rakhmawati, 2018).

The validation or analysis of learning assessment instruments is crucial for ensuring the quality of the items used (Sudaryono, Rahayu, & Margono, 2013). This analysis covers several key aspects: the difficulty level of the items, their ability to distinguish between high- and low-performing students, and the effectiveness of the incorrect answer options in multiple-choice questions. The results of this analysis can be used to improve the quality of the questions. Determining the difficulty level and discriminative power of questions can be done through various methods, either manually or using software. One tool frequently used to support this process is AnBuso version 8.0 software, which is designed to facilitate the analysis of assessment instruments.

AnBuso is a simple software program designed to help teachers analyze test items. The application is based on Microsoft Excel, making it more user-friendly. The advantages of AnBuso include an intuitive interface, high compatibility, ease of use, support for objective and essay questions, the ability to group students for remedial work, the presentation of results in the form of structured reports, as well as visual graphs, and it is free to use (Muhson, Lestari, Supriyanto, & Baroroh, 2014). Therefore, AnBuso is ideal for teachers due to its ease of use. In this context, the researchers analyzed test items on problem-solving skills and a self-efficacy questionnaire on statistics material.

Prior research

Assessment in mathematics education serves to determine student learning outcomes and plays a crucial role in fostering meaningful learning experiences from both cognitive and affective perspectives. High-quality assessment instruments are essential to ensure that students' abilities, particularly their mathematical problem-solving skills, can be measured accurately and objectively. Therefore, item analysis must be conducted to assess the quality of the instruments by testing validity, reliability, difficulty level, and discriminative power so that the instruments used are truly suitable for the learning evaluation process (Pardimin, Widodo, & Purwaningsih, 2017).

Problem-solving skills are one of the students' key competencies that they must possess in 21st-century mathematics education. These skills reflect the students' thought processes in understanding problems, determining solution strategies, implementing solutions, and verifying the results obtained (Mataheru, Laurens, Moma, & Sampulawa, 2019). In addition to helping students solve mathematical problems, problem-solving skills also play a role in developing critical thinking and decision-making skills in daily life (Ormonoy, 2022). However, these skills cannot be measured optimally if the instruments used are not of high quality. Instruments that

lack validity and reliability have the potential to produce inaccurate information regarding students' abilities (Azwar, 2019).

In addition to cognitive abilities, affective aspects such as self-efficacy also play a significant role in mathematics learning. Self-efficacy refers to students' belief in their own ability to complete tasks or tackle challenges in mathematics. Students with high self-efficacy tend to be more confident, persistent, and able to use problem-solving strategies effectively when facing math problems (Yousuf & Rajeswari, 2024). Conversely, students with low self-efficacy tend to give up easily and view math as a difficult subject (Putri & Prabawanto, 2019).

The quality of assessment instruments can also influence students' affective state during the learning process. Well-designed assessment instruments that align with learning indicators can provide a more objective assessment experience and help boost students' confidence in completing math tasks (Nortvedt & Buchholtz, 2018). Therefore, analyzing assessment instruments is crucial for measuring learning outcomes and can support the development of students' attitudes and beliefs regarding math learning.

Although there has been a considerable amount of research on item analysis and self-efficacy in mathematics learning, the two are generally studied separately. Item analysis focuses more on the quality of cognitive instruments, such as validity, reliability, difficulty level, and discriminative power, without linking them to students' affective aspects. Meanwhile, research on self-efficacy emphasizes the relationship between students' self-beliefs and learning outcomes, without examining the quality of the instruments used. This separation means that learning evaluations do not yet provide a complete picture of students' abilities and psychological conditions in learning mathematics. To objectively and data-drivenly confirm this trend, an analysis was conducted on relevant publications from the past ten years using Publish or Perish (PoP) and visualized through VOSviewer.

An initial literature search was conducted using PoP with the Scopus database for the years 2016–2026. The search was performed directly by combining the keywords "item analysis" AND "self-efficacy" AND "mathematics education," linked by Boolean operators, to ensure that the identified publications truly incorporated the integration of these two constructs within the context of mathematics education. The search results showed that this keyword combination yielded 42 publications. The data were then further analyzed using VOSviewer to map thematic connections and identify research gaps.

The visualization in Figure 1 shows that the dominant clusters are primarily centered on pedagogical themes such as "mathematics teacher" and "teaching practice," as well as the development of self-efficacy beliefs. Meanwhile, terms that directly represent analytical processes, such as "item analysis," do not emerge as major clusters or exhibit strong connections with self-efficacy. This indicated that research on self-efficacy in mathematics education was more frequently examined within the context of teaching practices and the teacher's role, rather than within an integrative framework that links it to the quality of measurement instruments.

Studies integrating the assessment of cognitive and affective instruments within a single analytical framework have also not yet showed systematic integration, particularly at the junior high school level and in statistics courses. The use of technology for item analysis, such as the AnBuso application, was already available but had not been widely utilized in research as part of learning evaluation. Analysis of self-efficacy questionnaires also still relies on statistical techniques tailored to the specific data type. This highlighted the need for a study that combined item analysis of problem-solving tests using AnBuso with self-efficacy questionnaire evaluation,

thereby providing a more comprehensive understanding of the quality of assessment instruments in junior high school statistics education. AnBuso was selected for this study because it can automatically calculate analysis parameters including difficulty level, discriminative power, validity, and reliability and generate systematic, easily interpretable outputs. Compared to manual analysis, which is prone to calculation errors and requires relatively more time, the use of analysis software offers efficiency and consistency in results. Furthermore, although other software such as ITEMAN can also be used for item analysis based on classical test theory, AnBuso was chosen because it was more practical, compatible with the needs of classroom-scale research, and suitable for the context of learning evaluation at the junior high school level.

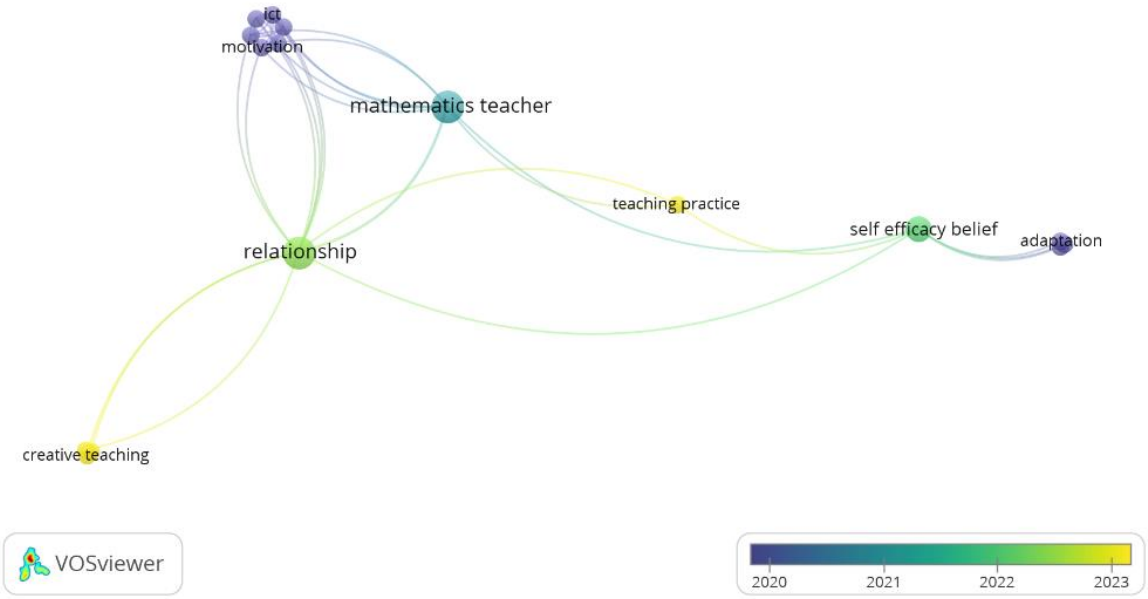


Figure 1. Keyword co-occurrence mapping of research on the integration of item analysis and self-efficacy

Methods

The research method used was quantitative descriptive to obtain data regarding the validity, reliability, difficulty level, and discriminative power of the test items. In addition, this study analyzed a self-efficacy questionnaire. This study focused on the eight-grade students at a public junior high school, SMP Negeri 1 Mlati. The research sample consisted of 32 students who are enrolled in statistics class. The instruments used in this study consisted of a problem-solving ability test and a self-efficacy questionnaire. This study limited its analysis to validity, reliability, discriminative power, and difficulty level.

Instrument

The research instrument used to collect data on problem-solving skills consisted of a test comprising 12 multiple-choice questions and 3 essay questions. The problem-solving skill indicators used in this study are presented in Table 1.

Table 1. Problem-solving skills indicators

No	Indicators	Description
1	Understanding the problem	Identifying whether there is sufficient data to solve the problem in order to gain a complete picture of what is known and what is being asked in the problem.
2	Planning the solution	Establishing the steps for the solution, selecting appropriate concepts, equations, and theories for each step.
3	Carrying out the plan	Carrying out the solution based on the steps that have been designed, using the selected concepts, equations, and theories.
4	Reviewing	Reviewing what has been done, verifying whether the solution steps were carried out as planned, and checking the accuracy of the answer to ultimately reach a final conclusion.

Next, a self-efficacy questionnaire for mathematics learning was used to collect data on students' self-efficacy. The questionnaire consisted of 20 items, each with five response options rated on a Likert scale: Always (A), Often (O), Sometimes (S), Rarely (R), and Never (N). In this study, the questionnaire was scored positively for positive statements and negatively for negative statements. The self-efficacy questionnaire indicators used are shown in Table 2.

Table 2. Student self-efficacy indicators

Indicators	Item number		Number of items
	(+)	(-)	
Students' confidence in learning and understanding mathematical concepts based on their abilities	1,2,3,4	5	5
Students' confidence in their ability to solve math problems or complete math-related assignments	6,7,8	9,10,11	6
Students' belief in their resilience when facing obstacles in learning mathematics	12,16	13,14,15	5
Students' confidence in their ability to achieve their mathematics learning objectives	17,18,19	20	4
Total	12	8	20

The problem-solving test items and self-efficacy questionnaire items used in this study were developed independently by the research team in accordance with the instrument development guidelines. Both instruments are available in full via this link.

Validity

Validity indicates the extent to which a test can accurately measure what it is intended to measure. Validity testing aims to assess the level of accuracy of the measurement instrument used (Saputri, Zulhijrah, Larasati, & Shaleh, 2023). In this study, the instrument validity was tested using content validity through expert judgment. The assessment was conducted by three validators, consisting of mathematics lecturers and teachers, using a 1–5 rating scale. The aspects assessed included the alignment of the material with the indicators, item construction, language use, and the instrument alignment with problem-solving skills and self-efficacy. The results of the validators' assessments were then analyzed using Aiken's V validity index, which is widely used in the analysis of instrument content validity (Azwar, 2019). The validity index is given by

$$V = \frac{\sum S}{[n(C-1)]}$$

where

$$\sum S = r - l_0$$

Description:
V = Aiken's validity coefficient
S = The score obtained from subtracting the lowest possible rating from the rating assigned by the expert
l_0 = The lowest rating score on the assessment scale
C = The highest rating score on the assessment scale
r = The rating assigned by the expert
n = The number of experts involved in the validation process

The Aiken index (V) ranges from 0 to 1. The validity categories used in this study are shown in Table 3.

Table 3. Validity categories

Aiken's validity index	Category
$\boldsymbol{V > 0.8}$	Highly valid
$\boldsymbol{0.4 < V \leq 0.8}$	Moderatly valid
$\boldsymbol{V \leq 0.4}$	Less valid

Source: Istiyono (2020)

Reliability

Reliability refers to the consistency of measurement results. An instrument is said to be reliable if it produces stable data across repeated measurements under the same conditions (Farida & Musyarofah, 2021). The reliability coefficient for multiple-choice items was calculated using the Kuder-Richardson method, known as KR-20, via the following equation.

$$KR-20 = \frac{K}{K-1}\left(1 - \frac{\sum P_i q_i}{S_{i^2}}\right)$$

Description:
K = Number of items
$P_i q_i$ = Variance of item scores
P_i = Proportion of correct responses to item i
q_i = Proportion of incorrect responses to item i
S_{i^2} = Total variance of respondents' scores

The reliability coefficient for the descriptive data can be calculated using the following Cronbach's alpha method:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right]$$

Description:
n = Number of test items
$\sum\sigma_i^2$ = Sum of the item variances
σ_t^2= Total variance of the test scores
The categories of reliability values used in this study are presented in Table 4.

Table 4. Reliability coefficient categories

Reliability coefficient	Category
$\boldsymbol{0.8 \leq \alpha \leq 1.0}$	Very high
$\boldsymbol{0.6 \leq \alpha < 0.8}$	High
$\boldsymbol{0.4 \leq \alpha < 0.6}$	Moderate
$\boldsymbol{0.2 \leq \alpha < 0.4}$	Low
$\boldsymbol{\alpha < 0.2}$	Very low

Source: Allen & Yen (1979)

Item discrimination index

Discriminatory power refers to the ability of a test item to distinguish between students with high and low ability. Put simply, the higher the discriminating power of a question, the more students from the higher-achieving group can answer correctly, whilst fewer students from the lower-achieving group can answer correctly (Purba, Fadhilaturrahmi, Purba, & Siahaan, 2021). The discriminating power of a question was determined using the following formula:

$$D = \frac{B_A - B_B}{N_A} \times 100\%$$

Description:
D = Item discrimination index
B_A= Number of students in the upper group who answered the item correctly
B_B= Number of students in the lower group who answered the item correctly
N = Number of students in either the upper group A or the lower group B
The discriminant power categories used in this study are presented in Table 5.

Table 5. Item discrimination categories

Item discrimination coefficient	Category
$\boldsymbol{D > 0.3}$	Good
$\boldsymbol{0.2 \leq D \leq 0.3}$	Fair
$\boldsymbol{D < 0.2}$	Poor

Source: Ebel & Frisbie (1991)

Item difficulty index

The level of difficulty refers to the degree of difficulty of a question, adjusted to the ability or level of knowledge of the individuals or group being assessed. The level of difficulty of a question was calculated using the following formula.

$$TK = \frac{B}{N} \times 100\%$$

Description:
TK = Item difficulty index
B = Number of students who answered the item correctly
N= Total number of students
The item difficulty criteria used in this study are presented in Table 6.

Table 6. Item difficulty categories

Item difficulty index	Category
$\boldsymbol{TK = 0.0}$	Very difficult
$\boldsymbol{0.0 < TK < 0.3}$	Difficult
$\boldsymbol{0.3 \leq TK \leq 0.7}$	Moderate
$\boldsymbol{0.7 < TK < 1.0}$	Easy
$\boldsymbol{TK = 1.0}$	Very easy

Source: Allen & Yen (1979)

Findings and Discussion

The results and discussion section of this study were organized into several subsections to help readers understand the content and findings of the study in a more structured manner. The discussion began with a presentation of the quantitative data obtained, followed by the results of the analysis carried out using AnBuso and Microsoft Excel software. Subsequently, an interpretation of these analytical results was provided to offer a deeper understanding. The data analyzed included scores from the test instruments and questionnaires used during the study. The presentation of this data aims to provide a clear picture of the quality of the test items and the level of self-efficacy of the students who were being subjects of the study. With this approach, the research results were explained systematically, starting from the presentation of raw data to the drawing of conclusions based on the results of the analysis carried out using relevant tools.

Table 7. Results of multiple-choice and essay tests analyzed using AnBuso

No	Participant	L/P	Objective test (60%)			Completion test (0%)	Essay test (40%)	Final score
			Correct	Incorrect	Score			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Respondent 1	P	8	4	66.67	0.00	90.00	76.00
2	Respondent 2	P	5	7	41.67	0.00	76.67	55.67
3	Respondent 3	P	7	5	58.33	0.00	93.33	72.33
4	Respondent 4	P	4	8	33.33	0.00	76.67	50.67
5	Respondent 5	P	5	7	41.67	0.00	73.33	54.33
6	Respondent 6	L	7	5	58.33	0.00	90.00	71.00
7	Respondent 7	L	8	4	66.67	0.00	86.67	74.67
8	Respondent 8	P	10	2	83.33	0.00	100.00	90.00
9	Respondent 9	L	7	5	58.33	0.00	93.33	72.33
10	Respondent 10	L	9	3	75.00	0.00	100.00	85.00
11	Respondent 11	P	5	7	41.67	0.00	73.33	54.33
12	Respondent 12	P	4	8	33.33	0.00	83.33	53.33

No	Participant	L/P	Objective test (60%)			Completion test (0%)	Essay test (40%)	Final score
			Correct	Incorrect	Score			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
13	Respondent 13	L	7	5	58.33	0.00	93.33	72.33
14	Respondent 14	P	8	4	66.67	0.00	93.33	77.33
15	Respondent 15	L	6	6	50.00	0.00	80.00	62.00
16	Respondent 16	L	7	5	58.33	0.00	93.33	72.33
17	Respondent 17	P	7	5	58.33	0.00	93.33	72.33
18	Respondent 18	L	8	4	66.67	0.00	90.00	76.00
19	Respondent 19	P	7	5	58.33	0.00	80.00	67.00
20	Respondent 20	P	10	2	83.33	0.00	96.67	88.67
21	Respondent 21	P	7	5	58.33	0.00	96.67	73.67
22	Respondent 22	L	8	4	66.67	0.00	80.00	72.00
23	Respondent 23	L	7	5	58.33	0.00	96.67	73.67
24	Respondent 24	L	9	3	75.00	0.00	96.67	83.67
25	Respondent 25	P	10	2	83.33	0.00	100.00	90.00
26	Respondent 26	P	11	1	91.67	0.00	96.67	93.67
27	Respondent 27	L	9	3	75.00	0.00	86.67	79.67
28	Respondent 28	L	11	1	91.67	0.00	93.33	92.33
29	Respondent 29	L	12	0	100.00	0.00	100.00	100.00
30	Respondent 30	L	11	1	91.67	0.00	90.00	91.00
31	Respondent 31	L	11	1	91.67	0.00	93.33	92.33
32	Respondent 32	P	12	0	100.00	0.00	96.67	98.67

Table 7 displays the analysis outcomes for problem-solving test items: there is a difference in student performance between objective (multiple-choice) and essay (open-ended) tasks. Generally, scores on essay questions tended to be higher than on objective questions. This indicated that some students could develop open-ended answers better than selecting the correct option in multiple-choice questions. Nevertheless, as objective tests carried a greater weighting, this component continued to have a strong influence on the final marks obtained by students.

On the other hand, the students' final marks were mostly in the medium to high range, although variations between the students were still evident. This variation indicated that the instrument used could distinguish the students' abilities to a certain extent. This finding provided a basis for further analysis of the quality of the test items, particularly regarding the level of difficulty and the discriminative power they generated.

Table 8. Self-efficacy questionnaire data

Respondent	Items																				Total score
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
AB	5	4	5	5	4	5	4	4	5	4	5	5	5	4	5	5	5	4	4	4	91
AN	4	4	4	4	3	4	4	4	4	3	4	4	5	4	4	4	4	4	4	3	78
ALA	5	4	4	4	3	4	4	5	5	4	5	4	4	4	4	4	5	4	4	4	84

Respondent	Items																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total score
ARA	4	4	4	5	4	5	5	5	5	4	5	5	4	4	4	4	4	4	4	4	87
AIM	5	4	5	4	4	4	4	4	5	4	4	4	5	4	4	4	4	4	5	5	86
BMF	5	5	5	5	5	5	5	4	5	5	5	5	5	4	5	5	5	5	5	5	98
DAI	4	5	4	4	5	5	4	4	4	4	5	4	4	4	5	4	4	4	4	4	85
DOW	4	4	4	4	4	4	4	4	4	3	5	5	4	4	4	5	5	3	4	3	81
FHAH	4	4	4	4	4	4	4	4	4	4	5	5	5	4	5	5	4	5	5	3	86
FRA	4	4	4	4	4	4	4	4	4	4	5	5	4	4	4	5	4	4	4	4	83
FAS	5	5	5	5	5	5	5	5	5	5	4	4	4	5	4	4	4	4	4	4	91
FF	5	5	5	5	5	5	5	4	5	5	4	4	4	4	4	4	4	4	4	4	89
GAP	5	5	5	5	5	4	4	4	4	5	4	5	5	5	4	5	4	4	4	4	90
HNA	5	5	5	5	4	5	4	5	5	4	5	5	4	5	5	5	5	5	5	5	96
HP	4	4	4	4	4	4	5	4	4	4	5	4	4	4	4	4	5	5	5	5	86
JMPS	5	4	4	4	4	4	4	5	5	4	4	4	4	5	5	5	4	4	5	4	87
JRA	5	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	5	4	3	4	80
KAA	4	4	4	4	4	4	3	4	4	4	4	4	5	4	4	4	4	4	4	4	80
K	4	4	5	4	4	4	4	4	4	4	5	5	5	5	4	5	4	4	4	3	85
LAW	5	5	4	4	4	4	4	4	4	5	5	5	5	5	4	5	4	4	4	4	88
MKI	4	4	5	5	4	5	5	5	5	5	4	4	4	5	5	5	5	5	5	5	94
MAF	5	4	5	4	4	5	4	4	5	5	4	4	4	4	4	5	4	4	3	3	84
MFK	4	5	5	4	4	4	4	5	5	4	5	5	4	5	4	5	4	4	4	4	88
MRA	5	4	4	4	4	4	3	4	4	4	5	4	4	4	5	5	5	5	5	5	87
NAZ	5	4	5	4	4	5	4	5	4	4	4	4	4	4	4	5	4	5	5	5	88
NH	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	80
OZP	4	4	4	4	4	4	4	5	5	4	4	5	4	4	4	5	5	4	5	5	87
RAS	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	100

Respondent	Items																				Total score
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
RAR	4	4	4	4	3	4	4	4	5	4	4	4	4	4	4	5	4	4	4	4	81
YSA	5	4	5	5	5	5	4	5	5	5	5	4	4	5	4	4	5	4	4	5	92
ZPA	4	4	4	4	4	4	4	4	4	4	5	4	4	4	4	4	5	4	4	4	82
ZAS	4	4	4	5	5	5	5	5	4	4	5	5	4	5	5	5	5	4	4	4	91
Mean total score																					87.03

Table 8 displays the analysis outcomes for questionnaire. There one can see the students' self-efficacy scores tend to fall within the high category. Most respondents scored above 80, with a maximum score of 100 and a minimum of 78. This distribution suggested that students generally possessed a high level of self-confidence when learning statistics.

The mean score of 87.03 indicated that students' self-efficacy was at a high level. However, there were still individual differences. This result suggested that students' levels of self-confidence were not entirely uniform. This variation warranted further examination in relation to the quality of the questionnaire instrument used, particularly through validity and reliability analyses to ensure that the data obtained truly reflected students' self-efficacy.

Analysis of problem-solving test items

The problem-solving research instrument consisted of 15 questions: 12 multiple-choice questions, each with a maximum score of 1 for a correct answer and 0 for an incorrect one, and 3 essay questions, each with a maximum score of 10. Before further analysis, the instrument was validated, and the validation results indicated a validity coefficient of 0.90 (very high) for the problem-solving ability test. Subsequently, its reliability was analyzed using Cronbach's alpha, calculated using Microsoft Excel, whilst the test's discriminant validity and level of difficulty were calculated using AnBuso. The results can be seen in Table 9, 10, 11, and 12, consecutively

Table 9. Reliability of multiple-choice items

Component	Value
Number of items	12
Sum of item variances	2.593
Total variance	4.843
Value $\boldsymbol{r_{11}}$	0.51
Category	Moderate

The outcomes in Table 9 showed that the reliability of the multiple-choice items was sufficient for measuring students' abilities. The total variance obtained indicated that there was variation in students' responses to the items.

Table 10. Reliability of the essay items

Component	Value
Number of items	3
Sum of item variances	2.983
Total variance	5.718
Value r_{11}	0.72
Category	High

The outcomes in Table 10 showed that the reliability of the open-ended tasks fell into the high category. It indicates that the instrument showed good consistency in measuring problem-solving ability. The total variance obtained was greater than that of the multiple-choice items, implying more significant variation in responses among students. Based on these reliability analysis results, it was confirmed that this problem-solving test instrument could be used effectively to measure problem-solving ability regarding statistical contents.

Table 11. Item discrimination and difficulty analysis of the multiple-choice items

Item No	Item discrimination		Item difficulty		Overall evaluation
	Coefficient	Category	Coefficient	Category	
(1)	(2)	(3)	(4)	(5)	(6)
1	-0.048	Poor	0.688	Moderate	Poor
2	0.308	Good	0.594	Moderate	Good
3	0.211	Fair	0.688	Moderate	Good
4	0.235	Fair	0.563	Moderate	Good
5	-0.009	Poor	0.594	Moderate	Poor
6	-0.116	Poor	0.656	Moderate	Poor
7	0.241	Fair	0.594	Moderate	Good
8	-0.018	Poor	0.688	Moderate	Poor
9	0.427	Good	0.688	Moderate	Good
10	0.490	Good	0.781	Easy	Fair
11	0.317	Good	0.688	Moderate	Good
12	0.498	Good	0.813	Easy	Fair

As seen in Table 11, the analysis outcomes of the discrimination indices for the 12 multiple-choice questions revealed that 41.67% fell into the 'good' category, 25% into the 'fairly good' category, and 33.33% into the 'poor' category. This distribution indicated that less than half of the items possessed optimal discriminative power, whilst the remaining third still required improvement due to low or negative discrimination coefficients. Items with negative discrimination indices indicated potential weaknesses in question construction, inaccuracies in the answer key, or distractors that did not function effectively.

In terms of difficulty level, 83.33% fell into the moderate category, and 16.67% were classified as easy, with no items in the difficult category. This composition was relatively proportionate for learning assessment, as the predominance of the moderate difficulty level allows the instrument to function reliably in measuring variations in students' abilities. However, the quality and clarity of some questions still needed to be improved so that the instrument could more effectively evaluate the students' problem-solving abilities regarding statistical material.

Table 12. Item discrimination and difficulty analysis of the multiple-choice items

Item No	Item discrimination		Item difficulty		Overall evaluation
	Coefficient	Category	Coefficient	Category	
(1)	(2)	(3)	(4)	(5)	(6)
1	0.536	Good	0.866	Easy	Fair
2	0.533	Good	0.888	Easy	Fair
3	0.569	Good	0.794	Easy	Fair

In the essay-type instrument, which consisted of three questions, 100% of the items showed good discriminatory power (see Table 12). In terms of difficulty level, 100% were classified as easy. Nevertheless, the high discriminatory power coefficients indicated that differences in the students' abilities could still be identified by the quality of their problem-solving strategies, procedural accuracy, and mathematical reasoning. These findings suggested that these questions were effective in assessing students' problem-solving abilities regarding statistical material.

Self-efficacy questionnaire analysis

The self-efficacy questionnaire was designed to identify the students' level of confidence in their ability to complete tasks or achieve specific goals in mathematics learning. The questionnaire consisted of 20 statements, comprising 12 positive and 8 negative items, and used a Likert scale to determine the scores. As with the previous problem-solving ability test, the questionnaire had been validated and found to be valid. The following are the reliability results of the questionnaire, calculated using Cronbach's Alpha formula in Microsoft Excel.

Table 13. Results of the reliability analysis of the student self-efficacy questionnaire

Components	Value
Number of items	20
Sum of item variances	23.647
Total variance	114.356
Value r_{11}	0.83
Category	High

Table 13 shows that the total variance for all items falls within the high category. High variance indicated that the student self-efficacy questionnaire had good internal consistency. This suggested that the measurement tool used was effective in assessing students' level of confidence in their ability to complete tasks or achieve specific goals in mathematics learning and could therefore be used for future evaluations.

Obviously, problem-solving ability in mathematics is rooted in systematic stages as outlined by Pólya (1945), namely understanding the problem, planning the solution, executing the plan, and reviewing the solution. This thinking structure confirms that the quality of a problem-solving instrument is determined by the level of difficulty of the questions and the extent to which the questions are able to distinguish the depth of students' thinking strategies. The findings of this study indicated that the predominance of items in the moderate category, along with discrimination indices largely falling within the fair to good categories, suggested that the instrument has been effective in identifying variations in students' strategic abilities. However, there were still items that required revision.

The findings of this study indicated that the quality of an instrument could not be assessed solely from a statistical perspective but must also be understood in relation to the theories of measurement, problem-solving, and self-efficacy in an integrated manner. From the perspective of Classical Test Theory, a good instrument is characterized by proportional difficulty levels, adequate discriminative power, and consistent reliability. The research results showed that most questions fell into the moderate difficulty category and had good discriminatory power. This indicated that the instrument could identify variations in student ability with reasonable stability. Research by Andersson and Wiberg (2017) also confirms that instruments with a predominance of moderate difficulty tend to be more effective in distinguishing student ability than instruments that are too easy or too difficult.

In relation to George Pólya's theory of problem-solving, the quality of test items is crucial because problem-solving ability measures not only the final outcome but also the students' thought processes in understanding the problem, planning strategies, carrying out the solution, and checking their work. Questions of moderate difficulty allow students to carry out Pólya's four stages optimally without immediately encountering failure or solving the problem in an overly simplistic manner. This aligns with the PISA framework developed by the OECD (2019), which states that mathematical problem-solving questions should provide opportunities to explore thinking strategies, rather than merely measuring procedural skills. Thus, the predominance of moderate-difficulty questions in this study indicates that the instrument sufficiently supports the measurement of problem-solving ability more authentically.

Furthermore, the presence of several items with negative or low discriminative power can be understood not only as a technical weakness of the instrument but also as an indicator that the questions have not yet fully succeeded in representing the stages of problem-solving thinking according to Pólya. Items with low discriminative power allow high- and low-ability students to provide relatively similar answer patterns, thereby reducing the instrument's ability to identify the quality of thinking strategies. According to Rodriguez (2005), this condition is often related to ineffective distractors, ambiguous wording, or a mismatch between the indicators and the question construction. Therefore, revisions to several items, particularly multiple-choice questions with negative or low discriminative power, are necessary so that the instrument is better able to capture differences in the quality of students' thinking processes when solving mathematical problems.

On the other hand, students' self-efficacy scores in the high category (mean 87.03) indicated that the students had a strong belief in their mathematical abilities. From the perspective of Bandura (1997), social cognitive theory, self-efficacy is formed through mastery experiences, social reinforcement, and meaningful learning experiences. When linked to Pólya's theory, the students' success in progressing through the stages of understanding a problem to arriving at a solution could serve as a mastery experience that reinforces their self-confidence in tackling subsequent mathematical tasks. In other words, problem-solving instruments that presented challenges at a moderate level of difficulty had the potential not only to measure cognitive ability but also to foster positive learning experiences that enhanced students' self-efficacy.

The relationship between the quality of problem-solving instruments and self-efficacy also indicated a connection between cognitive and affective aspects in mathematics learning. Instruments that are too difficult can undermine students' self-confidence by leading to repeated experiences of failure, whilst instruments that are too easy do not sufficiently stimulate the development of thinking strategies. Therefore, the composition of questions with a

predominance of moderate difficulty in this study was relevant not only from the perspective of measurement theory but also from the perspective of self-efficacy development. This finding was supported by the research of Talsma, Schüz, Schwarzer, and Norris (2018), which showed that self-efficacy had a positive relationship with mathematical achievement, particularly when students were engaged in structured and meaningful problem-solving activities.

The implications of these findings for mathematics learning were quite significant. Firstly, the composition of questions, with a predominance of moderate difficulty, can be maintained as a formative assessment strategy to obtain a stable picture of students' abilities. Secondly, items with low discriminative power need to be revised so that the instrument is truly capable of identifying differences in ability with greater precision. Thirdly, the high level of students' self-efficacy indicates that the instrument and the accompanying learning experiences have supported the development of academic self-confidence, which in the long term has the potential to enhance students' self-confidence in tackling more complex mathematics learning.

Although this study provides an empirical picture of the quality of the problem-solving instrument and the student self-efficacy questionnaire, there are several limitations that need to be considered when interpreting the results. Firstly, the sample size used was relatively limited, involving only 32 students from a single school at Year 8 level. This means the research findings cannot yet be widely generalized to a more diverse population with different school characteristics and student backgrounds. Secondly, the item quality analysis in this study still employed the Classical Test Theory approach. It means that the difficulty and discriminant power parameters are highly dependent on the characteristics of the sample used. This approach has not yet produced invariant item parameter estimates. Thirdly, the self-efficacy questionnaire instrument uses a self-report method based on a Likert scale, meaning that the measurement results are highly dependent on the students' subjective perceptions.

Conclusion

This study highlights the importance of analyzing problem-solving items and self-efficacy questionnaires in mathematics education, particularly in the context of statistics. The results of the analysis, conducted using AnBuso and Microsoft Excel, indicated that overall, the majority of the problem-solving test items had a good quality, with a good discriminant power of 53.33% and a moderate level of difficulty of 66.67%. This indicated that the test was sufficiently reliable for use in measuring students' mathematical problem-solving skills, although some test items still required improvement.

Furthermore, the self-efficacy questionnaire yielded positive results with good reliability. It indicated that the instrument could be used to measure students' level of confidence in their ability to complete tasks and achieve mathematical learning objectives. The instrument developed in this study could be used in the evaluation of mathematics learning with revisions to several test items, particularly multiple-choice questions that have negative or low discriminative power. However, revisions to these items are necessary to improve the quality, accuracy, and effectiveness of the assessment instrument.

This study provides insights for educators on improving the quality of assessment instruments used in learning. By taking these analysis results into account, it is hoped that teachers and curriculum developers can design better assessment tools to support effective and high-quality learning processes. Careful and systematic evaluation is essential to maintain educational standards and ensure that learning objectives are achieved optimally.

References

Alifia, N. N., & Rakhmawati, I. A. (2018). Kajian kemampuan self-efficacy matematis siswa dalam pemecahan masalah matematika. *Jurnal Pembelajaran Matematika*, *5*(1), 44–54. Retrieved from https://jurnal.uns.ac.id/jpm/article/view/26024

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Complete ed.).* New York, NY: Longman.

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika*, *82*(1), 48–66. https://doi.org/10.1007/s11336-016-9528-7

Arifin, Z. (2012). *Evaluasi pembelajaran*. Bandung: Remaja Rosdakarya.

Arikunto, S. (2019). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.

Astuti, A. P., Aziz, A., Sumarti, S. S., & Bharati, D. A. L. (2019). Preparing 21st century teachers: Implementation of 4C character's pre-service teacher through teaching practice. *Journal of Physics: Conference Series*, *1233*(1). https://doi.org/10.1088/1742-6596/1233/1/012109

Azwar, S. (2019). *Reliabilitas dan validitas (Edisi 4)*. Yogyakarta: Pustaka Pelajar.

Balqis, Patmawati, H., & Yulianto, E. (2024). Perbedaan kemampuan siswa dalam memecahkan masalah matematika berdasarkan tingkat self-confidence pada materi statistika. *Jurnal Kongruen*, *3*(2), 162–169. Retrieved from https://jurnal.unsil.ac.id/index.php/kongruen/article/view/12336

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman and Company.

Căprioară, D. (2015). Problem solving - purpose and means of learning mathematics in school. *Procedia - Social and Behavioral Sciences*, *191*, 1859–1864. https://doi.org/10.1016/j.sbspro.2015.04.332

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Farida, F., & Musyarofah, A. (2021). Validitas dan reliabilitas dalam analisis butir soal. *Al-Muarrib Journal of Arabic Education*, *1*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Florea, N. M., & Hurjui, E. (2015). Critical thinking in elementary school children. *Procedia - Social and Behavioral Sciences*, *180*, 565–572. https://doi.org/10.1016/j.sbspro.2015.02.161

Inoue, N., Asada, T., Maeda, N., & Nakamura, S. (2019). Deconstructing teacher expertise for inquiry-based teaching: Looking into consensus building pedagogy in Japanese classrooms. *Teaching and Teacher Education*, *77*, 366–377. https://doi.org/10.1016/j.tate.2018.10.016

Istiyono, E. (2020). *Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern*. Yogyakarta: UNY Press.

Kartowagiran, B. (2012). *Penulisan butir soal*. Yogyakarta: Universitas Negeri Yogyakarta.

Khotimah, N. H., Khoirunnisa, A., & Bilda, D. W. (2020). Pengaruh self-efficacy siswa SMP terhadap pemecahan masalah pada materi aritmetika sosial. *EDISI: Jurnal Edukasi Dan Sains*, *2*(2), 285–291. Retrieved from https://ejournal.stitpn.ac.id/index.php/edisi

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

Mataheru, W., Laurens, T., Moma, L., & Sampulawa, H. (2019). Analysis of student thinking processes in mathematical problem solving using saintificial approach. Proceedings of the 1st International Conference on Advanced Multidisciplinary Research (ICAMR 2018). Presented at the Proceedings of the 1st International Conference on Advanced Multidisciplinary Research (ICAMR 2018), Makassar, Indonesia. https://doi.org/10.2991/ICAMR-18.2019.51

Muhson, A., Lestari, B., Supriyanto, S., & Baroroh, K. (2014). Pengembangan software analisis butir soal yang praktis dan aplikatif. *Jurnal Ilmu Pendidikan Universitas Negeri Malang*, *20*(2), 110163. https://doi.org/10.17977/JIP.V20I2.4618

NCTM. (2000). *Standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics. *National Council of Teachers of Mathematics*, *7*(2), 1–16. Retrieved from https://books.google.com/books/about/Principles_and_Standards_for_School_Math.html?hl=id&id=BkoqAQAAMAAJ

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: Responding to issues regarding methodology, policy, and equity. *ZDM 2018 50:4*, *50*(4), 555–570. https://doi.org/10.1007/S11858-018-0963-Z

Nuutila, K., Tapola, A., Tuominen, H., Molnár, G., & Niemivirta, M. (2021). Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task

engagement. *Learning and Individual Differences*, *92*, 102090. https://doi.org/10.1016/j.lindif.2021.102090

OECD. (2019). *PISA 2018 results (Volume I) what students know and can do*. https://doi.org/10.1787/5f07c754-en

Okyranida, I. Y., Mayanty, S., & Widiyatun, F. (2024). Analisis butir soal kemampuan berpikir kritis siswa SMAIT Nururrohmah Depok. *Jurnal Penelitian Pembelajaran Fisika*, *15*(1), 73–79. https://doi.org/10.26877/jp2f.v15i1.17057

Ormonoy, T. (2022). The importance of solving math problems in elementary school: Pentingnya memecahkan masalah matematika di sekolah dasar. *Indonesian Journal of Education Methods Development*, *17*(4), 10.21070/ijemd.v20i.628-10.21070/ijemd.v20i.628. https://doi.org/10.21070/IJEMD.V20I.628

Pardimin, P., Widodo, S. A., & Purwaningsih, I. E. (2017). Analisis butir soal tes pemecahan masalah matematika. *WACANA AKADEMIKA: Majalah Ilmiah Kependidikan*, *1*(1). https://doi.org/10.30738/WA.V1I1.1084

Polya, G. (1945). *How to solve it*. Princeton, NJ Princeton University Press. Scientific Research Publishing.

Purba, Y. O., Fadhilaturrahmi, Purba, J. T., & Siahaan, K. W. A. (2021). *Teknik uji instrumen penelitian pendidikan*. Bandung: Widina Bhakti Persada.

Putri, W. K. H. W., & Prabawanto, S. (2019). The analysis of students' self-efficacy in learning mathematics. *Journal of Physics: Conference Series*, *1157*(3), 032113. https://doi.org/10.1088/1742-6596/1157/3/032113

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing.

Retnawati, H., Hadi, S., Nugraha, A. C., Ramadhan, M. T., Apino, E., Djidu, H., … Sulistyaningsih, E. (2017). *Menyusun laporan hasil asesmen pendidikan di sekolah*. Yogyakarta: UNY Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Saputri, H. A., Zulhijrah, Larasati, N. J., & Shaleh. (2023). Analisis instrumen assesmen: Validitas, reliabilitas, tingkat kesukaran dan daya beda butir soal. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, *9*(5), 2986–2995. https://doi.org/10.36989/DIDAKTIK.V9I5.2268

Sudaryono, Rahayu, Wardani., & Margono, Gaguk. (2013). *Pengembangan instrumen penelitian pendidikan*. 174. Yogyakarta: Graha Ilmu.

Sudijono, Anas. (2017). *Pengantar evaluasi pendidikan*. Jakarta: RajaGrafindo Persada.

Syaiful, S., Kamid, K., Muslim, M., & Huda, N. (2020). Identifiying of problem-solving abilities in mathematics among junior high school students. *Journal of Education and Learning (EduLearn)*, *14*(2), 176–182. https://doi.org/10.11591/edulearn.v14i2.14861

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, *61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Wong, T. T. Y., & Yip, E. S. K. (2023). What is the unknown? the ability to identify the semantic role of the unknown from word problems longitudinally predicts mathematical problem-solving performance. *Contemporary Educational Psychology*, *73*, 102183. https://doi.org/10.1016/j.cedpsych.2023.102183

Yousuf, Ms. I., & Rajeswari, G. (2024). Relationship between mathematics self-efficacy and mathematics achievement. *International Journal of Emerging Knowledge Studies*, *03*(10), 797–802. https://doi.org/10.70333/IJEKS-03-10-011