PTK AND EDUCATION
E-ISSN: 2549-2535 | P-ISSN: 2460-1780
Vol. 11 No. 02 (2025), pp. 149-156
DOI: 10.18592/ptk.v11i02.19385

EMPIRICAL VALIDITY OF SCIENCE ASSESSMENT INSTRUMENTS BASED ON PROBLEM SOLVING ON ELECTRICAL, MAGNETIC AND ENERGY MATERIALS

Didik Darmansyah[1], Ida Farida Chaidir[2], Ade Yeti Nuryantini[3]
[123]Universitas Islam Negeri Sunan Gunung Jati
ade.yeti@uinsgd.ac.id

Abstract
This research is motivated by the low problem-solving ability of students and the dominance of low-level cognitive instruments (LOTS) compiled by teachers, which often fail to accurately measure high reasoning power. The purpose of this study is to analyze the empirical validity of problem-solving-based science assessment instruments on electrical, magnetic, and alternative energy materials. The research method used was quantitative descriptive involving 32 students in grade IX. Data were collected through multiple-choice test instruments and analyzed using the Anates 4.0.9 program and SPSS to measure validity, reliability, difficulty level (ITK), and differentiating power (IDB). The results showed that 65% of the questions were declared empirically valid with a very high reliability coefficient of 0.870. The level of difficulty was dominated by the medium category (60%), but it was found that 35% of the questions had poor discriminating power. Qualitative findings from expert validators also reveal gaps in the synchronization between indicators and image stimuli in magnetic matter. It can be concluded that this instrument meets the requirements for strong measurement regularity, but requires improvements in the synchronization aspect of contextual stimuli to measure high-level thinking skills (HOTS). Overall, 55% of the questions are suitable for immediate use, 30% need revision, and 15% must be discarded.
Keywords:
Empirical Validity, Science Assessment Instruments, Problem Solving, Question Item Analysis

Abstrak
Penelitian ini dilatarbelakangi oleh rendahnya kemampuan pemecahan masalah siswa dan dominansi instrumen kognitif tingkat rendah (LOTS) yang disusun guru. Tujuan penelitian ini adalah menganalisis validitas empiris instrumen penilaian IPA berbasis pemecahan masalah pada materi listrik, magnet, dan energi. Metode deskriptif kuantitatif digunakan dengan subjek 32 siswa kelas IX. Instrumen 20 butir soal pilihan ganda dianalisis menggunakan program Anates 4.0.9. Hasil penelitian menunjukkan 65% soal valid secara empiris dengan reliabilitas sangat tinggi (0,870). Tingkat kesukaran didominasi kategori sedang (60%), namun 35% daya pembeda soal masih jelek. Berdasarkan integrasi data empiris dan *expert judgment*, disimpulkan 55% soal layak digunakan, 30% memerlukan revisi, dan 15% dibuang. Instrumen ini telah memenuhi target sebagai alat evaluasi yang reliabel, namun memerlukan sinkronisasi indikator dan stimulus soal yang lebih kuat untuk mengukur keterampilan berpikir tingkat tinggi (HOTS) secara optimal.
Kata Kunci:
Validitas Empiris, Instrumen Penilaian IPA, Pemecahan Masalah, Analisis Butir Soal

INTRODUCTION

Education is a continuous and planned process to produce quality human resources who have critical thinking and are able to face future challenges (Magdalena et al., 2021). In the 21st century, the demands of education graduates are not only limited to mastering theory, but must have 4C skills (*Communication, Collaboration, Critical thinking, dan Problem Solving*) (Saraswati et al., 2021). In the context of Natural Science (IPA) learning, problem-solving skills are at the heart of the learning process because it trains students to identify problems, design strategies, and draw conclusions based on scientific evidence from real-life phenomena (Lukman et al., 2023a).

Ideally, learning evaluation should be able to measure the achievement of educational goals comprehensively, covering high-level cognitive domains such as analyzing (C4), evaluating (C5), and creating (C6) (L. Masitoh & Aedi, 2020). Teachers as the main evaluators are expected to have competence in compiling valid, reliable assessment instruments, and have the right level of difficulty and differentiation so that learning outcomes can present students' true abilities (Maemonah, 2022). High-quality instruments are crucial to ensure the accuracy of measurement results and the accuracy of the data obtained (Saputri & Larasati, 2023).

The main problem in this study is the low quality of teacher-made evaluation instruments which are still dominated by the low-level cognitive realm (LOTS), such as the aspects of

remembering (C1) and understanding (C2), so that they fail to accurately measure students' problem-solving abilities (Lukman et al., 2023a). Although the demands of 21st century education require graduates to have analytical thinking and problem-solving (HOTS) skills, the reality on the ground shows that teachers often find it difficult to interpret the cognitive level of Bloom's Taxonomy into valid and reliable question items (L. Masitoh & Aedi, 2020). This condition is exacerbated by the results of international studies such as PISA and TIMSS which consistently rank Indonesian students' science and mathematics literacy skills at the bottom because of their unfamiliarity with non-routine problems that demand high reasoning (Yuliantaningrum & Sunarti, 2020). In abstract and complex materials such as electricity, magnetism, and energy, students often experience misconceptions because existing assessment instruments have not been able to diagnose the error of their thinking process in depth (Nurhayati, 2021).

Including the reality on the ground shows that there is a significant gap. The results of these international studies (PISA and TIMSS results) consistently place the science and mathematics literacy skills of Indonesian students at the bottom, which indicates the low ability of students to solve non-routine problems that require high reasoning (Lukman et al., 2023a). Many assessment instruments designed by teachers are still fixated on the low-level cognitive domain, especially the aspects of knowledge (C1) and comprehension (C2), and tend to take questions from package books without conducting an in-depth analysis of the quality of the question items (Alti et al., 2021). In addition, teachers often face difficulties in developing instrument-based *Higher Order Thinking Skills* (HOTS) because of the limited understanding of the rules for writing questions that are valid and empirically reliable (Alwahid dkk., 2023).

This condition can also be seen in abstract science materials such as electricity, magnetism, and energy, where students often experience misconceptions and difficulties in connecting concepts with their mathematical applications (Nurhayati, 2021a). Based on the results of the item analysis, some questions are declared valid, while others are invalid; Similarly, some have a good discriminating power index and some show a poor discriminating power index, thus demanding a decision to retain, revise, or discard the item (Muslim Darmawan et al., 2022). Empirical question item analysis activities are mandatory steps that must be taken to improve the quality of the instrument and minimize bias in measuring students' abilities (Setiyawan & Wijayanti, 2020).

This study aims to analyze the empirical validity of problem-solving-based science assessment instruments on electrical, magnetic, and energy materials, which includes validity, reliability, difficulty, and differentiating power (Akhmadi, 2021). The novelty of this research lies in the development of an instrument specifically designed to evaluate or diagnose students' ability to think processes in solving contextual problems that are real and familiar to students in complex physics material, beyond the assessment of the final result alone (Zahra et al., 2025). It is hoped that this research can make an innovative contribution in providing high-quality evaluation tools for science teachers and help familiarize students with analytical and systematic thinking according to the demands of the current global era (Setiyawan & Wijayanti, 2020).

METHODS

This study uses a quantitative descriptive approach (Akhmadi, 2021). The main objective is to describe and analyze the quality of test items empirically based on data obtained from field testing (Mutakin, 2023). In contrast to theoretical studies, this study focuses on the actual performance of the instrument when applied to a specific group of subjects to determine its feasibility in measuring student learning outcomes (Surya et al., 2023).

This research method is designed in detail using the *Research and Development* (R&D) with ADDIE's development model that includes the *Analysis, Design, Development, Implementation,* and *Evaluation* (Ardania et al., 2022). In the development stage, the instrument consisting of 20 multiple-choice questions and description questions is arranged based on a grid that integrates problem-solving indicators according to Polya or Brookhart (Kurniasi & Arsisari, 2020). The validity of the instrument was tested comprehensively through two tracks: the validity of the contents by the experts (*expert judgment*) includes expert validators (Tadris Science Lecturers) along with experienced practitioners, namely 2 science teacher colleagues, who function as examiners of material, construction, and language aspects, as well as empirical validity through field trials to 32 grade IX students (L. Masitoh & Aedi, 2020).

The test results in this study are analyzed using computer programs such as Anates 4.0.9 or can use other calculation applications such as SPSS to calculate the correlation coefficient *Product Moment* for validity, *Alpha Cronbach* for reliability, as well as the difficulty level index (ITK) and differentiating power (IDB) to ensure that each question item is able to objectively separate the groups of high- and low-ability students (Akhmadi, 2021). The materials used in this study are test items that include Electrical Materials (Static and Dynamic), Magnets, and Alternative Energy Sources. The research instrument includes multiple-choice questions containing 20 questions designed to evaluate problem-solving skills (Ardania et al., 2022). Even if it is in the form of an essay question, it only focuses on the conceptual and mathematical aspects of physics material (Nurhayati, 2021). In addition, validation sheets are also used for experts to assess the content, language, and suitability of instruments with cognitive levels (C3–C6) (Lukman et al., 2023a).

The data collection process follows the following systematic stages, including 1) The preparation stage, namely by designing a grid of questions based on the Independent Curriculum and in accordance with standards for problem solving (Ardania et al., 2022; Lukman et al., 2023b). 2) The expert assessment stage is carried out by means of the instrument undergoing a qualitative review by experts to assess its accuracy and scientific construction (Magdalena et al., 2020). 3) The field testing stage, namely with question items in validated instruments given to 32 student subjects using tests based on digital platforms such as *Smartschool* PGII and Anates for data collection or processing (Akhmadi, 2021). And the last one is 4) The documentation stage, which is by recording the students' responses/answers and converting them into raw scores for further statistical processing (Subakti, 2025).

The techniques and criteria for data analysis carried out are data collection which is then analyzed using Anates version 4.0.9 and Microsoft Excel which is used manually (using calculation formulas) to evaluate several specific quality criteria (Setiyawan & Wijayanti, 2020), such as performing Validity testing using product moment correlation. An item is considered empirically valid if the correlation coefficient (*Rx y*) greater than the value *Table* (0.514 for *N=32* at a significance level of 5%) (Akhmadi, 2021). As for the Reliability or reliability test using *Alpha Cronbach*. The instrument is considered Reliability (reliable) and consistent if the coefficient is≥0.70 (Subakti, 2025). Then measure the Difficulty Level Index (ITK) by determining the proportion of correct answers. Questions are classified as Easy (0.71–1.00), Medium (0.31–0.70), or Hard (0.00–0.30) (Surya et al., 2023). The Discrimination Index or Discriminating Power Index is by comparing the performance of the top 27% of students (upper group) and the bottom 27% of students (lower group) (L. F. Masitoh & Aedi, 2020). Items with an index ≥ value of 0.20 are considered "Excellent," while values below 0.20 are considered "Poor" and should be discarded or revised (Akhmadi, 2021). And the last is the effectiveness of the wrong answer choice or ditractor, that is, for multiple-choice questions, the wrong answer choice is considered effective if chosen by at least 5% of students (Akhmadi, 2021; Ciptasari et al., 2024; Saraswati et al., 2021).

RESULTS AND DISCUSSION

Results

The results of the data analysis carried out included testing the validity, reliability, level of difficulty, differentiation, and effectiveness of deception on 20 questions that had been tested on 32 students. Validity testing is carried out by comparing values *Rxy* with *The table is 0.514 at a significance level of 5%. A valid instrument shows that the item is able to measure what should be measured correctly* (Saputri & Larasati, 2023). A summary of the results of the validity test is presented in Table 1.

Table 1. Summary of Instrument Validity Test Results

Criteria	Number of Questions	Question Item Number
Valid	13	2, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 18
Invalid	7	1, 3, 4, 8, 10, 17, 19, 20

Based on Table 1, there are 65% of questions that are declared empirically valid. Invalid question items are caused by several factors, including unclear question construction or not in line with the desired problem-solving indicator.

Reliability tests are used to see the accuracy of the instrument in providing measurement results (Purniasari et al., 2021). The results of the calculation using the Alpha Cronbach formula

show the value of *R11=0 .870, which is in the category of very high reliability. This indicates that the instrument is very consistent and trustworthy to be used in learning evaluation* (Surya et al., 2023).

Difficulty level analysis aims to determine the proportion of the ease or difficulty of a question, while the differentiating power measures the ability of the question in distinguishing the upper and lower group of students (Saputri & Larasati, 2023). The distribution of the Difficulty Level and Differentiating Strength criteria is summarized in Table 2.

Table 2. Difficulty Level and Differentiating Analysis

Aspects	Criteria	Quantity	Percentage
Difficulty Level	Easy	6	30%
	Medium	12	60%
	Difficult	2	10%
Differentiating Power	Very good	6	30%
	Good	5	25%
	Enough	2	10%
	Ugly/Bad	7	35%

The data in Table 2 shows that this instrument is dominated by questions in the medium category (60%) (Mutakin, 2023), which is theoretically the ideal proportion for a learning outcome test. However, there are 35% of questions that have poor differentiating power, which are generally filled with a level of difficulty that is too easy to distinguish students' competencies significantly (Ningrum et al., 2023).

A distractor analysis of multiple-choice questions showed that most of the cheeter options worked well because they were chosen by at least 5% of test takers. However, in certain question items such as numbers 10 and 17, tricks are found that "don't work" or "badly" so that it makes it easier for students to guess the correct answer without a deep thinking process.

Based on the integration of all analysis criteria, the final decision on the use of the item as described in Table 3 is determined.

Table 3. Percentage of Decision Status Question Item

No.	Validity	Reliability	Difficulty Level	Differentiating Power	Offensive Power				Conclusion
					A	B	C	D	
1	Invalid	very high	Easy	Enough	Accepted	Accepted	Accepted	Revision	Revised
2	Valid	very high	Medium	Good	Accepted	Accepted	Accepted	Accepted	Worn
3	Invalid	very high	Medium	Ugly	Accepted	Accepted	Accepted	Accepted	Removed
4	Invalid	very high	Medium	Enough	Accepted	Accepted	Accepted	Rejected	Removed
5	Valid	very high	Medium	Good	Accepted	Accepted	Accepted	Revision	Worn
6	Valid	very high	Easy	Ugly	Accepted	Revision	Accepted	Rejected	Revised
7	Valid	very high	Medium	Good	Accepted	Accepted	Accepted	Revision	Worn
8	Invalid	very high	Easy	Ugly	Revision	Accepted	Accepted	Rejected	Revised
9	Valid	very high	Medium	Good	Accepted	Accepted	Accepted	Accepted	Worn
10	Invalid	very high	Difficult	Ugly	Accepted	Accepted	Accepted	Revision	Revised
11	Valid	very high	Medium	Good	Rejected	Accepted	Accepted	Accepted	Worn
12	Valid	very high	Medium	Enough	Accepted	Accepted	Accepted	Accepted	Worn
13	Valid	very high	Medium	Good	Accepted	Accepted	Rejected	Accepted	Worn
14	Valid	very high	Easy	Good	Rejected	Accepted	Accepted	Accepted	Worn
15	Valid	very high	Easy	Enough	Accepted	Rejected	Accepted	Revision	Worn
16	Valid	very high	Medium	Good	Rejected	Revision	Accepted	Accepted	Worn
17	Invalid	very high	Difficult	Ugly	Accepted	Accepted	Accepted	Accepted	Revised
18	Valid	very high	Easy	Enough	Accepted	Revision	Accepted	Accepted	Worn
19	Invalid	very high	Medium	Ugly	Accepted	Accepted	Accepted	Rejected	Removed
20	Invalid	very high	Easy	Ugly	Accepted	Accepted	Rejected	Revision	Revised

Table 3 informs that 55% of the questions (11 items) are suitable for immediate use, 30% (6 items) require revision, especially in the aspects of language and deception, while 15% (3 items) must be discarded because they do not meet the requirements for validity and discriminating power.

Discussion

The results of the analysis using the Alpha Cronbach formula show that this instrument has a reliability coefficient value of 0.870, which is statistically in the very high category (Saraswati et al., 2021). The achievement of this number gives a substantial meaning that the instrument has an extraordinary level of consistency, which means that if this measuring instrument is given back to the same subject at different times or occasions, then the measurement results will be relatively stable and consistent (Saputri & Larasati, 2023).

These findings show that the question items have been systematically arranged in accordance with the rules of good instrument writing, so as to minimize measurement errors (*measurement error*) that often happens in teacher-made tests (Akhmadi, 2021). Success in achieving a coefficient above 0.80 is also in line with the quality standards of the instrument-based *Higher Order Thinking Skills* (HOTS), which proves that even though the problem requires high reasoning and complex thinking processes, this instrument is still able to produce stable and reliable data (L. F. Masitoh & Aedi, 2020).

Comprehensively, this very high reliability is an absolute requirement for evaluation instruments to be reliable as a foundation for accountability and instructional decision-making by teachers (Akhmadi, 2021). Without strong reliability, teachers' decisions in mapping students' ability rankings (upper and lower groups) or determining learning completeness will be biased (Saputri & Larasati, 2023). Therefore, this instrument with a value of 0.870 has met the requirements for empirical validity to be used widely in evaluating the understanding of science concepts in electrical, magnetic, and energy materials accurately (Surya et al., 2023).

Empirically, 65% of the questions were declared valid. However, the results of qualitative judgment provide a critical perspective that is not captured by numbers. The validator, namely a fellow science teacher, found that question items 15 and 16 were not in harmony with the indicators due to the absence of images of magnetic or electromagnetic interactions in the existing questions. These findings confirm the theory that content validity through expert judgment is much more crucial than just statistical validity. KKO Incompatibility (*Operational Verbs*) in question number 1 (from C2 to C4) that the validator found showed that teachers are often stuck in the use of low verbs to measure high-level analytical skills. This is in accordance with the phenomenon in the field where teacher assessment instruments are still often dominated by the low-level cognitive realm (C1-C2) (Lukman et al., 2023b; Zahra et al., 2025).

This instrument was dominated by questions in the medium category (60%), followed by easy (30%) and difficult (10%) in Table 2. Theoretically, this proportion is close to the ideal normal curve (3-4-3 or 25-50-25) to optimize information regarding student ability variations (Ciptasari et al., 2024; Saraswati et al., 2021). However, it was found that 35% of the questions had poor differentiating power as shown in Table 2. In classical test theory, the low discriminating power of "easy" questions occurs because almost all students in the upper and lower groups can answer correctly, so that the questions lose their function to separate students who are truly competent from those who are less competent (Akhmadi, 2021; Ningrum et al., 2023). The implication is that these questions must be revised or discarded so as not to become biased in the final assessment (Purniasari et al., 2021; Saraswati et al., 2021).

Problems such as the grounding phenomenon on fuel trucks (Number 6) and the bimetal principle (Number 1) are innovative efforts to measure contextual problem-solving skills (Azizah et al., 2020). The data findings showed that students found it easier to identify problems (61.54%) than to evaluate solutions (30.77%) (Nurhayati, 2021b).

This reinforces Brookhart's theory that high-level thinking as problem-solving requires complex mental effort beyond just remembering facts (Hanafi et al., 2022; L. F. Masitoh & Aedi, 2020). The low ability of students to evaluate solutions is in line with the PISA results which shows the weaknesses of Indonesian students in the aspect of reasoning and concept integration in new situations (Jin et al., 2025; Zahra et al., 2025).

The implication of the research findings is that there is an increase in teacher competence, namely with the findings regarding the incompatibility of indicators and question items (on magnetism and energy problems) emphasizing the need for intensive training for teachers in compiling grids and instruments-based *Higher Order Thinking Skills* (HOTS) that is valid in substance or construction (Lestari et al., 2020; Pratiwi, 2021). Then the increase in the use of technology,

namely the use of applications such as Anates or SPSS, has been proven to make it easier for teachers to diagnose question items quickly, accurately, and informatively to determine which questions are worthy of entering the question bank (Akhmadi, 2021). And the last is the reflection of learning which is the result of an evaluation that shows that the dominance of easy questions requires teachers to modify learning strategies to be more challenging and applicative, so that students are used to facing non-routine challenges in accordance with the demands of the 21st century (Hanafi et al., 2022; Pratiwi, 2021).

CONCLUSION

This research succeeded in aligning the theoretical expectations contained in the introduction with the empirical reality in the field. Initially, this instrument was designed to answer the challenges of the 21st century that demand problem-solving skills (HOTS), in order to minimize reliance on routine memorization problems. The final results showed that this instrument has very strong psychometric qualities with a reliability of up to 0.870, which means that this measuring instrument is very steady and consistent for repeated use. However, in terms of validity, it is found that the statistical numbers do not always reflect the accuracy of the substance; Although 65% of the questions were declared empirically valid, *expert judgments* revealed that there was a "gap" in the synchronization between the image stimulants, indicators, and Operational Verbs (KKO) used.

An in-depth interpretation of these findings shows that students' abilities are still stuck at the problem identification stage and are not optimal at the solution evaluation stage. This confirms that the transition from knowledge-based instruments (LOTS) to problem-solving (HOTS) requires a high level of construction precision, not just increasing the difficulty of the problem. Overall, with 55% of questions that are suitable for immediate use and 30% that require revision, this instrument has met its target as a valid and reliable evaluation tool reference for complex physics materials.

As for suggestions and developments for the future, especially for Educators and Practitioners, it is necessary to make continuous adjustments between the achievement indicators and the KKO used (for example, changing the KKO 'identifying' to 'analyzing') to be in line with the high-level cognitive targets (C4-C6). For instrument development, it is recommended to add more representative visual illustrations, especially in the case of magnetic and transformer materials, to help students visualize contextual problems more realistically. And for further research, the potential for the development of this research lies in testing the validity of constructs and predictive validity with a wider and more diverse sample. In addition, the use of technology such as Large Language Models (LLM) in automatically diagnosing students' thinking processes can be an innovative research direction in the future to provide *more granular and faster* feedback.

THANKS

The author expressed his gratitude for the completion of the research on the empirical validity of this science assessment instrument as part of efforts to improve the quality of educational evaluation. The highest gratitude and appreciation are conveyed to all lecturers of the Science Education Evaluation course, who have provided guidance, constructive input, and direction in compiling research instruments to meet the rules of writing valid and reliable questions. The author also thanks the Tadris Sains Study Program of UIN Sunan Gunung Djati for facilitating a learning place for the author to develop pedagogic competencies, especially in designing assessments based on *Higher Order Thinking Skills* (HOTS).

Sincere appreciation is expressed to the PGII 1 Bandung Junior High School institution, especially to the Principal and teacher staff, who have allowed and assisted the author in the process of collecting field data. Thank you to 32 grade IX students of SMP PGII 1 Bandung who have been willing to be research subjects, so that empirical data related to the level of difficulty and differentiating power of question items can be obtained accurately. The author also thanks the expert validators who have provided qualitative assessments and suggestions for improvements to the material, construction, and language aspects of this instrument. Finally, gratitude is addressed to the families and fellow students who have provided moral support, motivation, and technical

assistance during this study. Hopefully all the assistance provided will be righteous deeds for everyone. Amen....

REFERENCES

Akhmadi, M. (2021). Analisis butir soal evaluasi tema 1 kelas 4 sdn plumbungan menggunakan program anates. *Ed-Humanistics: Jurnal Ilmu Pendidikan*, *Query date: 2025-12-19 10:02:30*. https://scholar.archive.org/work/6pvgfbdhmbcc7mycqwxpc2zplq/access/wayback/http://ejournal.unhasy.ac.id/index.php/ed-humanistics/article/download/1464/1048

Alti, R., Lufri, L., Helendra, H., & ... (2021). Instrumen asesmen berbasis literasi sains tentang materi keanekaragaman hayati. *Journal for Lesson and …*, *Query date: 2025-12-19 10:02:30*. https://ejournal.undiksha.ac.id/index.php/JLLS/article/view/34270

Alwahid, M., Khairuddin, K., Sepriadi, S., & Febrian, M. (2023). Evaluasi Analisis Instrumen Pembelajaran Kognitif Mata Pelajaran Pendidikan Jasmani Olahraga dan Kesehatan. *Jurnal JPDO*, *Query date: 2025-12-19 10:02:30*. http://jpdo.ppj.unp.ac.id/index.php/jpdo/article/view/1513

Ardania, R., Fitriyah, L., & ... (2022). Pengembangan Instrumen Soal Pilihan Ganda Berbantu Aplikasi Quizizz Materi Sistem Pernapasan Pada Manusia Untuk Peserta Didik Kelas VIII SMP. *Discovery: Jurnal Ilmu …*, *Query date: 2025-12-19 10:02:30*. https://pdfs.semanticscholar.org/7d82/f70768bb4136ca029fc21e81169ae951f659.pdf

Azizah, Z., Taqwa, M., & ... (2020). Analisis pemahaman konsep fisika peserta didik menggunakan instrumen berbantukan quizizz. … *Pendidikan …*, *Query date: 2025-12-19 10:02:30*. https://e-journal.iain-palangkaraya.ac.id/index.php/edusains/article/view/1707

Ciptasari, A. N., Sumarno, S., Wasikin H., E. H., & Risno Setiyono, R. S. (2024). ANALISIS BUTIR SOAL TES OBJEKTIF PENILAIAN TENGAH SEMESTER MATA PELAJARAN EVOLUSI DAN BIOTEKNOLOGI KELAS XII SMA. *EDUPROXIMA : Jurnal Ilmiah Pendidikan IPA*, *6*(2), 393–402. https://doi.org/10.29100/.v6i2.4488

Hanafi, M., Syamsuri, S., & ... (2022). Pengembangan instrumen soal higher order thinking skills (hots) matematika berdasarkan brookhart konteks motif batik pandegelang pada siswa mts. *Media Pendidikan …*, *Query date: 2025-12-19 10:02:30*. https://ojspanel.undikma.ac.id/index.php/jmpm/article/view/5207

Jin, H., Kim, Y., Jung, D., Kim, S., Choi, K., Son, J., & Kim, J. (2025). *Investigating Large Language Models in Diagnosing Students' Cognitive Skills in Math Problem-solving* (No. arXiv:2504.00843). arXiv. https://doi.org/10.48550/arXiv.2504.00843

Lestari, N., Hadiprayitno, G., Muhlis, M., & ... (2020). Pelatihan Teknik-Teknik Analisis Instrumen Penilaian Ranah Kognitif Guru SMPN 21 Mataram. … *Indonesian Journal Of …*, *Query date: 2025-12-19 10:02:30*. http://jpfis.unram.ac.id/index.php/jpmsi/article/view/8

Lukman, H. S., Setiani, A., & Agustiani, N. (2023). Pengembangan Instrumen Tes Kemampuan Pemecahan Masalah Matematis Berdasarkan Teori Krulik dan Rudnick: Analisis Validitas Konten. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *7*(1), 326–339. https://doi.org/10.31004/cendekia.v7i1.1761

Maemonah, M. (2022). Analisis Instrument Tes Multiple Choice Sebagai Alat Evaluasi Mata Pelajaran Ski Kelas Ix Di Mts Pringgabaya. … *Education Journals (Jurnal Ke-SD-An …*, *Query date: 2025-12-19 10:02:30*. https://ejournal.uniramalang.ac.id/primed/article/view/1363

Magdalena, I., Hifziyah, M., Aeni, V., & Rahayu, R. (2020). Pengembangan Instrumen Tes Siswa Tingkat Sekolah Dasar Kabupaten Tangerang. *Nusantara*, *Query date: 2025-12-19 10:02:30*. https://ejournal.stitpn.ac.id/index.php/nusantara/article/view/808

Magdalena, I., Syariah, E., Mahromiyati, M., & ... (2021). Analisis instrumen tes sebagai alat evaluasi pada mata pelajaran sbdp siswa kelas ii sdn duri kosambi 06 pagi. …, *Query date: 2025-12-19 10:02:30*. https://ejournal.stitpn.ac.id/index.php/nusantara/article/view/1264

Masitoh, L. F., & Aedi, W. G. (2020). Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) Matematika di SMP Kelas VII. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *4*(2), 886–897. https://doi.org/10.31004/cendekia.v4i2.328

Muslim Darmawan, -, S., Dwi Riyanti, Yohanes Gatot Sutapa Yuliana, & Sumarni. (2022). Test-Items Analysis of English Teacher-Made Test. *Journal of English Education and Teaching*, *6*(4), 498–513. https://doi.org/10.33369/jeet.6.4.498-513

Mutakin, D. (2023). Pemanfaatan Software SPSS untuk Analisis Instrument Penelitian Dalam Penelitian Tindakan Kelas. *Jurnal Inovasi Dan Manajemen Pendidikan*, *Query date: 2025-12-19 10:02:30*. https://pdfs.semanticscholar.org/a259/6c6a9a66750ba478bfebf73aabd49392ab3b.pdf

Ningrum, W. A., Rahmawati, R. C., Minarti, I. B., & Mekar, R. M. (2023). Analisis Butir Soal Ulangan Harian Pembelajaran IPA Pada Kelas VIII Di SMPN 21 Semarang. *Educatio*, *18*(1), 91–101. https://doi.org/10.29408/edc.v18i1.18291

Nurhayati, T. (2021). ANALISIS KEMAMPUAN PROBLEM SOLVING KONSEP FISIKA PADA MATERI IMPULS MOMENTUM PADA SISWA SMA MUHAMMADIYAH 1 DEMAK. *EduFisika: Jurnal Pendidikan Fisika*, *6*(2), 123–128. https://doi.org/10.59052/edufisika.v6i2.13512

Pratiwi, A. (2021). Penyusunan Instrumen Evaluasi Pembelajaran Berbasis Thinking Analysis Dalam Upaya Peningkatan Kompetensi Guru Matematika. *InEJ: Indonesian Engagement Journal*, *Query date: 2025-12-19 10:02:30*. https://jurnal.iainponorogo.ac.id/index.php/inej/article/view/2876/1757

Purniasari, L., Masykuri, M., & Ariani, S. (2021). Analisis butir soal ujian sekolah mata pelajaran kimia sma n 1 kutowinangun tahun pelajaran 2019/2020 menggunakan model iteman dan Rasch. *Jurnal Pendidikan Kimia*, *Query date: 2025-12-19 10:02:30*. https://jurnal.uns.ac.id/JPKim/article/view/48244

Saputri, H., & Larasati, N. (2023). Analisis Instrumen Assesmen: Validitas, Reliabilitas, Tingkat Kesukaran Dan Daya Beda Butir Soal. *Didaktik: Jurnal Ilmiah …*, *Query date: 2025-12-19 10:02:30*. http://journal.stkipsubang.ac.id/index.php/didaktik/article/view/2268

Saraswati, S., Rodliyah, I., & Rahmawati, N. (2021). Analisis Instrumen Penilaian Berbasis Higher Order Thinking Skills pada Mata Kuliah Matematika Lanjut. *Inomatika*, *Query date: 2025-12-19 10:02:30*. https://inomatika.unmuhbabel.ac.id/index.php/inomatika/article/view/275

Setiyawan, R., & Wijayanti, P. (2020). Analisis kualitas instrumen untuk mengukur kemampuan pemecahan masalah siswa selama pembelajaran daring di masa pandemi. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika …*, *Query date: 2025-12-19 10:02:30*. https://scholar.archive.org/work/px3abyrsljfrris2bls2gflgfy/access/wayback/http://lebesgue.lppmbinabangsa.id/index.php/home/article/download/26/16

Subakti, I. W. (2025). Analisis Kualitas Butir Soal Pada Uji Coba Evaluasi Pembelajaran Matematika. *CENDEKIA: Jurnal Ilmu Pengetahuan*, *Query date: 2025-12-19 10:02:30*. https://jurnalp4i.com/index.php/cendekia/article/view/4097

Surya, A., Sumarno, S., & Muhtarom, M. (2023). Analisis Kualitas Instrumen Tes Hasil Belajar IPAS Materi Wujud Zat dan Perubahannya. *Fondatia*, *Query date: 2025-12-19 10:02:30*. https://ejournal.stitpn.ac.id/index.php/fondatia/article/view/3190

Zahra, N. I. A., Handayani, L., & Isnaeni, W. (2025). *ANALISIS KEBUTUHAN PENGEMBANGAN INSTRUMEN TES KEMAMPUAN PEMECAHAN MASALAH MATEMATIS PESERTA DIDIK*. *10*.