Jurnal Evaluasi dan Pembelajaran
Volume 8 Nomor 1, Tahun 2026
Available online at https://jepjurnal.stkipalitb.ac.id/index.php/hepi

Psychometric Quality Analysis of Grade 5 Elementary School Mathematics Test Instruments: A Review Based on Validity, Reliability, and Difficulty Level Distribution Parameters

Yunni Arnidha[1*], Novalia[1], Undang Rosidin[1], Rabiyatul Adawiyah Siregar[1]
[1]Doctoral Program in Education, Faculty of Teacher Training and Education,
Universitas Lampung, Jalan ZA Pagar Alam No. 41 Gedungmeneng, Indonesia
**Corresponding author:* faizehsan@fkip.unila.ac.id

Abstract

This study aimed to analyze the psychometric quality of Grade 5 elementary school mathematics test instruments based on validity, reliability, and difficulty level distribution parameters. A quantitative descriptive approach was employed, with 15 multiple-choice items administered to 49 students as the unit of analysis. Data were collected through documentation of test scripts, answer keys, and student answer sheets. Analysis was conducted using Classical Test Theory, comprising item validity through item-total correlation, reliability using Cronbach's Alpha coefficient, and difficulty level analysis based on the proportion of correct answers. Results indicated that 9 items (60%) met validity criteria, while 6 items (40%) did not. The instrument obtained a Cronbach's Alpha value of 0.753, classifying it as reliable. Difficulty level distribution showed 8 items (53.33%) in the moderate category and 7 items (46.67%) in the easy category, with no items classified as difficult. The mathematics test instrument analyzed demonstrated adequate psychometric quality; however, it still requires revision of invalid items and development of items with more varied difficulty levels to ensure more accurate and comprehensive measurement of student ability.

Keywords: psychometric quality, test instrument, validity, reliability, difficulty level

Introduction

Education serves as a fundamental pillar in producing competitive human resources in the global era, where the quality of the national education system serves as a barometer of national progress. The 1945 Constitution Article 31 paragraphs (3) and (4) mandates the government to organize a quality education system integrated with the improvement of faith and noble character. This is reinforced in Law Number 20 of 2003 on the National Education System, which emphasizes that the learning process must holistically develop students' potential across cognitive, emotional, and psychomotor domains (Boini et al., 2023). In realizing this vision, learning evaluation occupies a strategic position as a systematic instrument for measuring the achievement of instructional objectives and providing feedback mechanisms for educators to undertake continuous improvement (Rossmalia et al., 2024).

In the discipline of mathematics, the quality of evaluation is particularly crucial because mathematics is a universal language that is foundational for everyday life and serves as the primary basis for mastery of science at higher educational levels (Zakiyah & Kartika, 2024). Theoretically, a high-quality evaluation instrument must fulfill psychometric parameters encompassing validity, reliability, level of difficulty, and discrimination power (Umiyati et al., 2026; Triani et al., 2023).

However, field realities indicate a significant disparity between theoretical standards and the implementation of evaluation in elementary schools. Many test instruments used in school examinations are developed without adequate empirical trials to ensure their psychometric validity (Boini et al., 2023). Educators are often trapped in using conventionally available test items without considering whether those items remain valid or require revision based on current student characteristics (Triani et al., 2023). Empirical evidence shows that mathematics items at the elementary school level often have disproportionate difficulty

distributions dominated by the 'easy' category thereby insufficiently exploring students' cognitive potential in depth (Umiyati et al., 2026).

The research gap lies in the inconsistency between the urgency of item analysis frequently emphasized in academic literature and the minimal practice of such scientific analysis at the elementary school level. Although Classical Test Theory (CTT) methodology has long been established, many schools still rely on static question banks without data-based updates (Umiyati et al., 2026; Syaifudin, 2023). The scarcity of research that comprehensively integrates validity, reliability, and difficulty distribution analysis for Grade 5 mathematics instruments creates an information gap regarding the precision of measurement tools used to map students' mathematical foundations.

The fundamental factor causing this gap is the administrative and technical burden perceived by educators, where the process of statistical item analysis is considered complex and time-consuming (Susanto et al., 2015). Limited mastery of educational evaluation technology also constitutes a major barrier for teachers in producing quality test items (Triani et al., 2023). The novelty of this study lies in its comprehensive psychometric quality review specifically focused on Grade 5 elementary school mathematics test instruments, highlighting precise difficulty level distribution analysis using contemporary software tools.

This study aimed to comprehensively describe the psychometric quality of mathematics test instruments used in Grade 5 elementary school. The primary objective was to analyze the characteristics of each item based on validity, reliability, and difficulty level distribution parameters. Through this empirical analysis, it is expected that items classified as good, requiring revision, or needing to be discarded can be identified to optimize the learning achievement assessment process.

Method

This study employed a quantitative descriptive research approach. Quantitative descriptive research aims to examine phenomena occurring at present by objectively and systematically describing and summarizing the characteristics of the object being studied (Ika Isna Umiyati et al., 2026; Susongko, 2016). The justification for using this research type is that the researcher intended to describe the psychometric quality of the mathematics test instrument that had been administered without imposing any treatment (Sulistyawati et al., 2022). The unit of analysis was the multiple-choice test items and the answer sheets of Grade 5 elementary school students (Setiawan et al. 2020).

The primary concept is psychometric quality, defined as the extent to which a test instrument is suitable for use as an accurate measuring tool for learning outcomes. The research variables analyzed included item characteristics: validity, reliability, and difficulty level (Retnawati, 2016). The operationalization of these concepts is presented in Table 1.

Table 1. Operationalization of Concepts, Variables, and Research Indicators

Concept	Variable	Indicator
Psychometric Quality of Test Instrument	Validity	The accuracy of test items in measuring the competence being tested through item-total score correlation (Boini et al., 2023).
	Reliability	The degree of consistency and stability of measurement results even when administered repeatedly (Rossmalia et al., 2024).

Concept	Variable	Indicator
	Level of Difficulty	The proportion of test participants who answered each item correctly(Umiyati et al., 2026).

The population consisted of all Grade 5 student answer sheets at UPT SDN 1 Patoman on the mid-semester mathematics examination for the even semester. Samples were determined using purposive sampling, a technique based on data representativeness and specific research objectives (Rossmalia et al., 2024; Setiawan et al. 2020). The total sample of 49 students was selected by considering the availability of student answer sheets and the requirements for statistical item analysis.

The data collection technique was documentation, obtaining secondary data comprising mathematics test scripts, item grids, answer keys, and student answer sheets processed by the school objectives (Rossmalia et al., 2024; Setiawan et al. 2020). This ensured that analyzed data represented actual evaluation results from the learning process (Novalia & Syazali, 2014).

For item validity testing of multiple-choice items, the Product Moment correlation formula was applied (Boini et al., 2023; Umiyati et al., 2026). An item was declared 'valid' if r-calculated > r-table at a 5% significance level. If the correlation was less than 0.30, the item was classified as invalid (Boini et al., 2023; Susongko, 2016).

Instrument reliability testing utilized the Cronbach's Alpha coefficient (Umiyati et al., 2026; Rossmalia et al., 2024). An instrument was declared 'reliable' if the reliability coefficient (r_{11}) > 0.60. Instruments with coefficients above 0.70 can be categorized as having good internal consistency and are suitable for use in learning evaluation (Basuki & Hariyanto, 2014; Boini et al., 2023).

The study was conducted at UPT SDN 1 Patoman, selected because the Grade 5 mathematics test instruments used there had never undergone empirical psychometric analysis, making standardization of the school's question bank necessary to ensure assessment fairness (Setiawan et al. 2020). Data analysis was performed using Classical Test Theory assisted by ANATES version 4.09 or IBM SPSS Statistics (Novalia & Syazali, 2014; Rossmalia et al., 2024). The difficulty index formula P = B/JS was used, where B is the number of students who answered correctly and JS is the total number of test participants.

Results and Discussion

Item Validity Analysis

Item validity was analyzed using Corrected Item-Total Correlation values, with the validity criterion of a correlation value greater than r-table of 0.282 (n=49, α=5%). The results indicated that not all items met empirical validity criteria, as presented in Table 2.

Table 2. Distribution of Item Validity

Category	Number of Items	Percentage
Valid	9	60.00%
Invalid	6	40.00%
Total	15	100%

The results showed that the majority of test items met validity requirements. However, six items had correlation coefficients below the minimum threshold and thus failed to

optimally represent the mathematical ability construct. The highest correlation value was found in item number 12 (r = 0.580), while the lowest was in item number 15 (r = -0.075). A negative correlation value indicates that the item is not functioning properly—students with high total scores tended to answer this item incorrectly, suggesting that this item needs to be revised or eliminated entirely.

This finding is consistent with Ndiung & Jediut (2020), who developed a mathematics learning achievement test instrument for elementary school students oriented toward higher-order thinking. Their study found that not all developed items met empirical validity criteria, necessitating revision before the instrument could be widely used. These findings confirm that empirical validation is a crucial stage in instrument development to ensure each item genuinely measures the targeted competency. Similarly, Yuliyanto et al. (2021) demonstrated that elementary school students' initial mathematical ability instruments require validity and reliability testing to be genuinely suitable as measurement tools.

In an international context, El-Hamamsy et al. (2022) affirmed the importance of validity and reliability in the development of learning assessment instruments; the Computational Thinking instrument they developed demonstrated strong content and construct validity along with adequate reliability, making it suitable for evaluating competency at the elementary school level. Based on these findings, the mathematics test instrument analyzed has adequate validity, but revision of the six invalid items is still needed to improve measurement precision.

Instrument Reliability Analysis

Instrument reliability was analyzed using Cronbach's Alpha coefficient. The results are presented in Table 3.

Table 3. Reliability Test Results

Indicator	Value
Number of Items	15
Cronbach's Alpha	0.753

The instrument obtained a Cronbach's Alpha value of 0.753, indicating a good level of internal consistency. This value demonstrates that the items within the instrument have adequate interrelatedness in measuring the mathematical ability construct of elementary school students. Consequently, the mathematics test instrument can be categorized as reliable and capable of producing relatively consistent scores when used with groups of participants with similar characteristics.

This finding is consistent with Pandra et al. (2017), who developed a mathematics achievement test for third-grade elementary school students and obtained a reliability coefficient of 0.783. Busnawir et al. (2023) also explained that the reliability coefficient is an important aspect in determining test quality, as stable reliability indicates that the instrument is capable of producing relatively consistent scores. El-Hamamsy et al. (2022) further confirmed that good reliability enables instruments to produce consistent estimates of learner ability across various sample groups. Therefore, the reliability value of 0.753 obtained in this study indicates that the instrument has met recommended reliability standards in educational research.

Difficulty Level Distribution Analysis

Difficulty level analysis was conducted to determine the proportion of students capable of answering each item correctly. The results are presented in Table 4.

Table 4. Difficulty Level Distribution

Category	Number of Items	Percentage
Difficult	0	0.00%
Moderate	8	53.33%
Easy	7	46.67%
Total	15	100%

The majority of items were in the moderate category (53.33%), while the remainder were in the easy category (46.67%). No items were classified as difficult. The highest difficulty index was found in item number 7 (P = 0.959), indicating a very easy item that almost all students answered correctly, while the lowest was in item number 15 (P = 0.326), approaching the moderate category boundary.

This finding is consistent with Hartono et al. (2023), who analyzed elementary school mathematics diagnostic test instruments using Classical Test Theory and found that item characteristics, including difficulty level, need to be analyzed to ensure instrument quality. Pandra et al. (2017) also emphasized that difficulty level analysis is important so that mathematics test instruments accurately describe variations in student ability. Their study concluded that moderate-category items have the best ability to describe student ability variations because they can be answered proportionally by both high and low ability students.

The absence of difficult items is a notable weakness of this instrument. Hartono et al. (2023) showed that evaluation instruments dominated by moderate-category items tend to produce more stable measurement information compared to instruments dominated by easy or difficult items, but also emphasized the importance of difficult items for identifying high-ability students. Argianti & Retnawati (2020) demonstrated that mathematics items dominated by easy categories reflect an imbalanced difficulty distribution, potentially reducing instrument quality in mapping student ability comprehensively.

Theoretically, Brookhart & Nitko (2018) explain that item analysis must consider difficulty and discrimination indices because both play a role in determining item quality and the instrument's ability to differentiate students by ability level. Therefore, the absence of difficult-category items means this instrument has not fully captured higher-order thinking abilities that are important in 21st-century mathematics learning. Subsequent instrument development needs to incorporate items with greater complexity so that the range of student ability can be measured more comprehensively.

Overall, the mathematics test instrument analyzed demonstrated adequate psychometric quality, as evidenced by the dominance of valid items, a reliability coefficient meeting standards, and a difficulty distribution dominated by moderate and easy categories. However, the six invalid items and absence of difficult items indicate that the instrument still requires revision and further development. This is consistent with Argianti & Retnawati (2020), who showed that imbalanced difficulty distribution can affect the instrument's information quality in mapping student ability. Instrument quality is not determined by one measurement aspect alone, but by the integration of validity, reliability, and item characteristics (Brookhart & Nitko, 2018).

Conclusions

Based on the data analysis results, this study concludes that the mathematics test instrument analyzed demonstrated adequate psychometric quality, though still requiring several improvements. First, validity analysis showed that 9 out of 15 items (60%) met validity criteria, while 6 items (40%) did not meet empirical validity criteria. This indicates that the

majority of items have been able to measure the mathematical ability construct targeted for measurement, though several items require revision for optimal functioning.

Second, reliability analysis showed that the instrument obtained a Cronbach's Alpha coefficient of 0.753, demonstrating good internal consistency and meeting reliability standards for use as a mathematics learning achievement evaluation tool. Third, difficulty level analysis showed that 53.33% of items were in the moderate category and 46.67% in the easy category, with no difficult-category items found. This condition indicates that the instrument is more dominant in measuring basic and intermediate student abilities, but has not fully identified students with higher-level mathematical abilities.

Overall, the Grade 5 elementary school mathematics test instrument analyzed can be declared sufficiently suitable for use as a learning evaluation tool. However, revision of invalid items and development of items with more varied difficulty levels should be conducted to improve the instrument's psychometric quality for future use.

Based on these findings, several recommendations are proposed. For educators, it is recommended to conduct periodic psychometric analysis of test instruments using Classical Test Theory-based software such as ANATES or IBM SPSS to ensure evaluation quality. The six invalid items should be revised by reviewing item language, clarity of answer choices, and alignment with learning indicators. Additionally, the development of difficult-category items is urgently needed to improve students' higher-order thinking skills. For future researchers, it is recommended to expand the scope of analysis to include discrimination power (daya beda) and distractor effectiveness to provide a more comprehensive picture of instrument quality.

References

Argianti, R., & Retnawati, H. (2020). Item analysis of mathematics multiple choice test in junior high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(2). https://scholarhub.uny.ac.id/jpep/vol24/iss2/4/

Basuki, I., & Hariyanto. (2014). *Asesmen pembelajaran*. PT Remaja Rosdakarya.

Boini, Asnawi, A., & Tisna, M. (2023). Analisis instrumen tes penilaian Pendidikan Agama Islam menggunakan tes teori klasik. *At Turots: Jurnal Pendidikan Islam*, *5*(1), 391–396. https://journal.stitmadani.ac.id/index.php/JPI/article/view/445

Brookhart, S. M., & Nitko, A. J. (2018). *Educational assessment of students* (8th ed.). Pearson. https://www.pearson.com/en-us/subject-catalog/p/educational-assessment-of-students/P200000000835/9780136976691

Busnawir, Kodirun, Zamsir, Samparadja, H., & Hasnawati. (2023). Trends stability of reliability coefficient based on sample size and ability of test-takers. *Pythagoras: Jurnal Matematika Dan Pendidikan Matematika*, *18*(2). https://doi.org/10.21831/pythagoras.v18i2.59392

El-Hamamsy, L., Zapata-Cáceres, M., Barroso, E. M., Mondada, F., Zufferey, J. D., & Bruno, B. (2022). The Competent Computational Thinking Test: Development and Validation of an Unplugged Computational Thinking Test for Upper Primary School. *Journal of Educational Computing Research*, *60*(7), 1818–1866. https://doi.org/10.1177/07356331221081753

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2023). Exploration of diagnostic testing instruments: Validity, reliability, and item characteristics. *Pegem Journal of Education and Instruction*. https://www.pegegog.net/index.php/pegegog/article/view/2147

Heri Retnawati. (2016). www.nuhamedika.gu.ma

Ika Isna Umiyati, Fina Fakhriyah, & Sumaji Sumaji. (2026). Analisis Kualitas Butir Soal Konsep Sistem Gerak Manusia Kelas VI Sekolah Dasar Berdasarkan Validitas, Reliabilitas, Tingkat Kesukaran dan Daya Beda. *Polygon : Jurnal Ilmu Komputer Dan Ilmu Pengetahuan Alam*, *4*(1), 51–64. https://doi.org/10.62383/polygon.v4i1.927

Ndiung, S., & Jediut, M. (2020). Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(1), 94–111. https://doi.org/10.25273/pe.v10i1.6274

Novalia, & Syazali, M. (2014). *Olah data penelitian pendidikan*. AURA.

Pamungkas, A., Dewantari, T., Tohir, A., ., S., & Soraya, R. (2026). The Effectiveness of the Flipped Classroom Model in Improving English Reading Comprehension of Seventh-Grade Junior High School Students. Journal of Cultural Analysis and Social Change, 11(1), 999–1006. https://doi.org/10.64753/jcasc.v11i1.4009

Pamungkas, A.,& Tohir, A. (2026). Evaluating the Effectiveness of Technology-Based Learning through an Evidence-Based Education Framework. (2026). EduTech Journal, 2(3), 1-12. https://doi.org/10.62872/4zkfbq42

Pandra, V., Sugiman, & Mardapi, D. (2017). Development of mathematics achievement test for third grade students at elementary school in Indonesia. *International Electronic Journal of Mathematics Education*, *12*(3), 769–776. https://iejme.com/download/development-of-mathematics-achievement-test-for-third-grade-students-at-elementary-school-in.pdf

Putra, H. S., Tohir, A., & Mashari, A. (2025). Analysis of Teacher' Digital Literacy Competencies as Support for Teachers' Skills at SDN 7 Gunung Agung. Jurnal Evaluasi dan Pembelajaran, 7(2), 206–216. https://doi.org/10.52647/jep.v7i2.400

Rossmalia, A. L., Zahra Wildaniar Amraow, G., Hakim, L., Pratiwi, V., & Negeri Surabaya, U. (2024). *Analisis Butir Soal Pilihan Ganda pada Elemen Akuntansi Lembaga atau Instansi Pemerintah Menggunakan Program Komputer Anates* (Vol. 4, Number 1).

Setiawan, M. A., Susongko, P., & Hayati, M. N. (n.d.). Pendeteksian DIF pada Perangkat Tes Objektif Penilaian Akhir Semester IPA dengan Mengunakan Permodelan Rasch. In *PSEJ* (Vol. 5). Retrieved http://scienceedujournal.org/index.php/psej

Singagerda, F.S., Tohir, A., Timotius, E., Fibrila, F., Asad, I., Rahim, R. (2026). A TOPSIS-Based Decision Support System for Selecting the Best Online Learning Platform. In: Bhateja, V., Hoong, A.L.S., Kong, J.D., Jena, J.J. (eds) Smart Computing Paradigms: Advanced Data Mining and Analytics. SCI 2025. Lecture Notes in Networks and Systems, vol 1683. Springer, Cham. https://doi.org/10.1007/978-3-032-08243-5_4

Sulistyawati, W., Wahyudi, & Trinuryono, S. (2022). Analisis (deskriptif kuantitatif) motivasi belajar siswa dengan model blended learning di masa pandemi COVID-19. *Kadikma*, *13*(1), 68–73.

Susanto, H., Rinaldi, A., & Islam Negeri Raden Intan Lampung, U. (2015). Analisis Validitas Reabilitas Tingkat Kesukaran dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. In *Jurnal Pendidikan Matematika* (Vol. 6, Number 2).

Susongko, P. (2016). *Pengantar metodologi penelitian pendidikan*. Universitas Pancasakti Tegal.

Syaifudin. (2023). Matematika dengan Item And Test Analysis. *Jurnal Jendela Matematika*, *1*(2).

Triani, L., Marlina, L., & Riak, S. (2023). *jurnalucy,+ANALISIS+BUTIR+SOAL+ULANGAN+HARIAN+KEMAGNETAN+DAN+INDUKSI+ELEKTROMAGNETIK+KELAS+IX+SMP*.

Yuliyanto, A., Turmudi, Syaodih, E., Saputra, D. R., Dharmawan, A., & Pertiwi, C. K. (2021). The validity and reliability of instruments for measuring elementary school students' early mathematical ability. *Kalamatika: Jurnal Pendidikan Matematika*, *6*(2), 127–142. https://doi.org/10.22236/KALAMATIKA.vol6no2.2021pp127-142

Zakiyah, Z., & Kartika, H. (2024). UJI VALIDITAS KONTEN INSTRUMEN KEMAMPUAN REPRESENTASI MATEMATIS DALAM MENYELESAIKAN MASALAH BANGUN DATAR. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, *10*(1), 250–257. https://doi.org/10.29100/jp2m.v10i1.5485