ISSN Online 2527-5933
MATH-EDU: Jurnal Ilmu Pendidikan Matematika
Volume 11, Nomor 1, April 2026, pp. 87-94

Quality Analysis of Evaluation Instruments for Junior High School Students' Mathematical Conceptual Understanding

Abdur Rahman Hakim[1]*, Rena Revita[2], Firza Mufti Aulia[3], Ade Irma[4]
[1,2,3,4] Faculty of Education and Teacher Training, State Islamic University of Sultan Syarif Kasim Riau, Riau, Indonesia.
*Correspondence: hakimduri2018@gmail.com

Article Information

Received:
11th August 2025

Accepted:
25th May 2026

Publish:
30th April 2026

Keywords
Evaluation instrument
Validity
Reliability
Discrimination power
Difficulty level

Abstract

This study investigates the quality of evaluation instruments developed to measure junior high school students' mathematical conceptual understanding, with a focus on the topic of numbers. The research employed a quantitative descriptive method involving thirty-two seventh-grade students from a junior high school in Pekanbaru. Data were obtained through essay-based assessments consisting of three items aligned with specific indicators of mathematical conceptual understanding and cognitive domain levels. The analysis examined validity, reliability, difficulty level, and discrimination power. The findings show that all three items achieved high validity, although the overall reliability coefficient was categorized as low. The difficulty level varied, with one item classified as easy and two as moderate. Discrimination power also differed among items, with one item having poor discrimination, one adequate, and one good. These results highlight the importance of conducting systematic quality testing to ensure that evaluation instruments are valid, reliable, and appropriately calibrated for assessing students' conceptual understanding in mathematics.

How to Cite: Hakim, A. R., Revita. R., Aulia, F. M., & Irma, A. (2026). Quality Analysis of Evaluation Instruments for Junior High School Students' Mathematical Conceptual Understanding. *Math-Edu: Jurnal Ilmu Pendidikan Matematika*, 11 (1), 87-94.

Introduction

Mathematical Conceptual Understanding (MCU) is an essential competence for students in the process of learning mathematics. This skill enables learners to comprehend mathematical concepts in a deeper and more meaningful way. Students who have strong MCU are able to interpret, explain, and draw conclusions about mathematical concepts based on their existing knowledge (Manul et al., 2019). According to Shadiq, as cited in Rismen et al., (2021), the indicators of MCU include the ability to restate a concept, the capacity to classify objects according to their attributes, the ability to provide examples and non-examples of a concept, the skill to present a concept in various mathematical representations, the ability to formulate necessary or sufficient conditions for a concept, and the capability to apply concepts in problem-solving.

The measurement of MCU can be carried out through evaluation. Evaluation serves as a tool to determine the extent to which an educational program, instruction, or training has achieved its objectives. The results of such evaluation can motivate both teachers and students to enhance learning engagement and improve the quality of their reasoning. Evaluation involves the process of systematically measuring and assessing the achievements obtained (Mania, 2014). Testing is one of the main methods for evaluating the learning process. Its purpose is to measure students' knowledge,

skills, and ability to solve problems. In designing test instruments, it is important to follow established criteria so that the test can function effectively as a tool for measuring students' abilities. A high-quality test instrument should meet the standards of evaluation which include validity, reliability, discrimination power, and an appropriate level of difficulty (Iskandar & Rizal, 2018).

Validity refers to the extent to which the collected data accurately represents what is intended to be measured (Sugiyono, 2021). A test is considered valid if it accurately measures the intended construct and produces results that are consistent with predetermined criteria (Solichin, 2017). Reliability, in contrast, refers to the consistency of an instrument in measuring the same variable. An instrument is reliable if it produces consistent results when administered to the same subjects under different conditions, times, or evaluators (Lestari & Yudhanegara, 2018).

Discrimination power indicates the ability of a test item to differentiate between high-performing and low-performing students. Items with high discrimination power can clearly distinguish between these groups of students (Wulan & Rusdiana, 2015). On the other hand, items with low discrimination power are considered less effective in differentiating students' abilities (Anwar et al., 2021).

The difficulty level of a test is another important factor that reflects how challenging an item is for students. A well-designed test contains items with a balanced level of difficulty, meaning that the items are neither too easy nor too difficult (Kania, 2019). Items that are too easy are ineffective in distinguishing between students with high and low abilities, while items that are too difficult may prevent even high-ability students from answering correctly (Hanifah, 2014)

The feasibility test of an evaluation instrument is crucial to ensure its effectiveness. However, many teachers give little attention to conducting such a test. This lack of attention can lead to the use of inadequate evaluation instruments, which may result in inaccurate or unfair assessments. For instance, a capable student might be assessed as having low competence, or the opposite might occur. Therefore, it is essential to examine whether the evaluation instrument being used meets the necessary standards (Nuraeni et al., 2015).

Previous research has shown that testing the quality of evaluation instruments is important to ensure their appropriateness in measuring students' learning outcomes. Loka (2019) found that a test instrument used to assess MCU in algebraic operations had high validity and moderate reliability. The first question was categorized as easy, the second and third as moderately difficult, and the fourth as difficult. Similarly, Rahmawati et al., (2013) discovered that an MCU test in trigonometry had varied levels of validity, ranging from low to high, and moderate reliability. The first three items had adequate discrimination power, while the other three items had good discrimination power. In terms of difficulty, the first three items were easy, while the remaining items were difficult.

Based on previous studies, it can be observed that the quality testing of instruments often focuses solely on mathematical problem-solving skills. Therefore, this study focuses on measuring a

different mathematical ability, namely MCU. The formulas used in this study also differ from those in previous research. It is expected that the findings of this study will contribute to the development of valid, reliable, and well-calibrated evaluation instruments with appropriate levels of difficulty and discrimination power. Such instruments can be used as references for designing more effective and targeted learning assessments, particularly for number topics at the junior high school level.

Methods

The purpose of this study is to analyze the quality of evaluation instruments for assessing students' Mathematical Conceptual Understanding (MCU) in the topic of numbers. This research employed a quantitative descriptive method with the participants consisting of thirty-two seventh-grade students from a junior high school in Pekanbaru. The data collection technique used was an essay test. The evaluation instrument consisted of three essay questions designed in accordance with the MCU indicators and the cognitive domain levels (C4 to C6), as shown in Table 1.

Table 1. Alignment of Questions with MCU Indicators

No.	Mathematical Ability Indicators	Cognitive Domain Level	Question
1	Applying or using concepts accurately in various situations.	C4	A minimarket has three refrigerators, each with a different rate of temperature decrease. The first refrigerator decreases by 4°C every 30 minutes, the second decreases by 1°C every 10 minutes, and the third decreases by 3°C every 15 minutes. If the initial temperature of each is 0°C, predict the order from the coldest to the warmest after one hour!
2	Restating a learned concept in one's own words.	C5	Given integers e, f, g, and h where $e > f > g > h$, prove that $(e + f)$ is always greater than $(g + h)$!
3	Presenting a concept in various mathematical representations	C6	Andi is playing by the Mahakam River. He sees a dolphin leaping 4 meters above the water surface and then diving 9 meters below the surface. Represent the dolphin's movement on a number line and determine the difference between the highest leap and the deepest dive!

After the questions were developed based on the MCU indicators and cognitive domain levels, they were validated and reviewed by lecturers and teachers to obtain feedback. Once the questions were deemed appropriate, a trial was conducted to assess the quality of the instrument. This trial included an item analysis to evaluate validity, reliability, difficulty level, and discrimination power.

Validity Testing

The validity of the instrument was measured using the Pearson product-moment correlation formula with raw scores as developed by Karl Pearson (Surapranata, 2009). The significance of the correlation

was tested using the t-test. If the calculated t-value was greater than the table value at a significance level of 5 percent, the item was considered valid. The validity coefficient was interpreted according to specific ranges, with values between 0,800 and 1.000 indicating very high validity, between 0,600 and 0,799 indicating high validity, between 0,400 and 0,599 indicating moderate validity, between 0,200 and 0,399 indicating low validity, and below 0,200 indicating very low validity.

Reliability Testing

Reliability was tested using the split-half method. First, item reliability was calculated using the Pearson product-moment formula. Then, the reliability of the entire test was determined using the Spearman-Brown formula (Arikunto, 2021). The reliability coefficient was interpreted as follows: 0,00 to 0,20 very low, 0,20 to 0,40 low, 0,40 to 0,60 moderate, 0,60 to 0,80 high, and 0,80 to 1.00 very high (Sudayana, 2015).

Difficulty Level Testing

The difficulty level was calculated using the formula from (Prijowuntato, 2016). Items were categorized as very difficult (0,00 to 0,20), moderately difficult (0,30 to 0,70), or easy (0,70 to 1.00) (Surapranata, 2009).

Discrimination Power Testing

The discrimination power of each item was calculated using the formula from(Lestari & Yudhanegara, 2018). Scores were interpreted as follows: ≤0,00 very poor, 0,00 to 0,20 poor, 0,20 to 0,40 adequate, 0,40 to 0,70 good, and 0,70 to 1.00 very good (Surapranata, 2009).

Result and Discussion

Result

The final semester evaluation to measure conceptual understanding was analyzed from several aspects, including validity, reliability, difficulty level, and discrimination power. The evaluation instrument consisted of three essay questions, each targeting different MCU indicators. Questions one and two had a maximum score of thirty-three, while question three had a maximum score of thirty-four, making the total score for all items one hundred. The analysis results are as follows.

Validity Analysis Results

Table 2. Validity Test Results

Question Number	t-value	t-table	Status	criteria
1	2,489	2,0422	Valid	Moderate
2	6,851	2,0422	Valid	High
3	5,400	2,0422	Valid	High

Based on Table 2, all three questions have t-values greater than the t-table value, indicating that they are valid. The validity level is predominantly high, with only question one categorized as moderate.

Reliability Analysis Results

The reliability coefficient obtained using the Pearson product-moment correlation was 0,2206. Applying the Spearman-Brown formula resulted in a coefficient of 0,3615. Comparing this with the r-table value at a significance level of 0,05 and degrees of freedom of thirty (0,3494), the instrument is considered reliable, although the reliability interpretation falls into the low category.

Difficulty Level Analysis Results

Table 3. Difficulty Level Test Results

Question Number	Difficulty Level	Criteria
1	0,64	Moderate
2	0,75	Easy
3	0,39	Moderate

As shown in Table 3, two questions have a moderate difficulty level, and one question, number two, is considered easy.

Discrimination Power Analysis Results

Table 4. Discrimination Power Test Results

Question Number	Discrimination Power	Criteria
1	0,168	Poor
2	0,405	Good
3	0,331	Adequate

From Table 4, it is evident that the discrimination power varies among the questions. Question two shows good discrimination power, question three is adequate, and question one is poor.

Discussion

Based on the conducted calculations, it can be concluded that the quality of each item, including its validity, reliability, difficulty level, and discrimination power, has been determined. For item one, the results indicate good validity with a moderate interpretation. This suggests that the question accurately measures the intended construct. This finding is consistent with the theory stating that a test is considered valid if its results align with predetermined criteria (Solichin, 2017). Although the item demonstrates reliability, the interpretation falls within the low category, indicating that the question is not yet consistent in producing the same results when administered under different testing conditions. The same situation applies to items two and three. Regarding discrimination power, item one is categorized as poor, which means it is not sufficiently capable of distinguishing between respondents with high and low abilities. This is in line with the theory that questions with low

discrimination power are less effective in differentiating between high-achieving and low-achieving students (Anwar et al., 2021). However, in terms of difficulty, item one is considered moderate, indicating that the level of challenge is reasonable.

For item two, the analysis reveals high validity. Nevertheless, its difficulty level is classified as easy, suggesting that the question is less effective in differentiating between respondents of different ability levels. This aligns with the opinion that questions with very easy difficulty levels are unable to clearly distinguish between high-ability and low-ability students (Hanifah, 2014). On the other hand, this item shows good discrimination power, indicating that it is effective in differentiating respondents with varying abilities. This finding supports the theory that questions with high discrimination power can clearly distinguish between students of different performance levels (Wulan & Rusdiana, 2015).

For item three, the results indicate high validity. Its discrimination power is categorized as adequate, meaning the item is effective in differentiating between high-performing and low-performing students, although improvements could be made. The difficulty level is classified as moderate, which suggests that the level of challenge is balanced. This finding is consistent with the view that a good question should have a proportional difficulty level, being neither too easy nor too difficult (Kania, 2019).

Conclusion

This study concludes that the three evaluated questions have very good validity levels, as indicated by t-values exceeding the t-table values. The instrument also demonstrates reliability, although the reliability interpretation falls into the low category. The difficulty levels of the questions vary, with question two being easy and questions one and three being moderate. Discrimination power also varies, with question one being poor, question three adequate, and question two good. This study was conducted in a single junior high school with specific student characteristics, so the results cannot be generalized to all schools. Future research should involve other schools with different conditions and student characteristics to strengthen the findings. In addition, the instrument only measured cognitive aspects of MCU. To gain a more comprehensive understanding of students' mathematical abilities, future instruments should also assess affective and psychomotor domains. Further studies could also explore factors influencing the quality of evaluation instruments, such as teacher competence, teaching materials, and school policies.

Recommendation

This study, which focused on a single junior high school with specific student characteristics, has limitations in the generalization of its findings. Therefore, replication in other junior high schools with different contexts and student characteristics is necessary to strengthen the results. Furthermore, the evaluation instrument employed in this research measured students' mathematical conceptual

understanding only in the cognitive domain. The development of a more comprehensive evaluation instrument, which also assesses the affective and psychomotor domains, is essential to provide a more complete picture of students' mathematical abilities. In addition, this study did not examine the factors influencing the quality of the evaluation instrument. Future research could explore aspects such as teacher competence, learning materials, and school policies to gain a deeper understanding of how to develop effective mathematics evaluation instruments.

References

Anwar, A. R. M., Siagian, T. A., Yensy, B., & Nurul, A. (2021). Analisis Kualitas Soal Matematika Penilaian Akhir Semester Ganjil Kelas VIII SMP Negeri 11 Kota Bengkulu. *Jurnal Penelitian Pembelajaran Matematika Sekolah (JP2MS)*, *5*(2), 267–280. https://doi.org/10.33369/jp2ms.5.2.267-280

Arikunto, S. (2021). *Dasar-dasar Evaluasi Pendidikan* (3 ed.). Bumi Aksara.

Hanifah, N. (2014). Perbandingan Tingkat Kesukaran, Daya Pembeda Butir Soal dan Reliabilitas Tes Bentuk Pilihan Ganda Biasa dan Pilihan Ganda Asosiasi Mata Pelajaran Ekonomi. *SOSIO e-KONS*, *6*(1), 41–55. https://doi.org/10.30998/sosioekons.v6i1.1715

Iskandar, A., & Rizal, M. (2018). Analisis Kualitas Soal di Perguruan Tinggi Berbasis Aplikasi TAP. *Jurnal Penelitian dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Kania, N. (2019). Kualitas Alat Evaluasi Hasil Belajar Matematika (Quality Of Evaluation Tools Of Mathematical Learning Results). *Jurnal Theorems*, *3*(2), 105–113. https://doi.org/10.31949/th.v3i2.1184

Lestari, K. E., & Yudhanegara, M. R. (2018). *Penelitian Pendidikan Matematika*. Refika Aditama.

Loka, S. (2019). Instrumentasi Kemampuan Pemecahan Masalah Matematis: Analisis Reliabilitas, Validitas, Tingkat Kesukaran Dan Daya Beda Butir Soal. *Gema Wiralodra*, *10*(1), 41–52. https://doi.org/10.31943/gemawiralodra.v10i1.8

Mania, S. (2014). *Asesmen Autentik untuk Pembelajaran Aktif dan Kreatif (Implementasi Kurikulum 2013)*. Alauddin University Press.

Manul, M. G., Susilo, D. A., & Fayeldi, T. (2019). Analisis Kemampuan Pemahaman Konsep Matematis Siswa Dalam Menyelesaikan Soal SPLDV Kelas X. *RAINSTEK : Jurnal Terapan Sains & Teknologi*, *1*(4), 45–53. https://doi.org/10.21067/jtst.v1i4.3655

Nuraeni, L., Elshap, D. S., & Kartika, P. (2015). Implementasi Penyusunan Instrumen Evaluasi Yang Digunakan Oleh Widyaiswara Dalam Mengukur Keberhasilan Pelatihan Di Balai Besar Pendidikan Dan Pelatihan Kesejahteraan Sosial Lembang. *Jurnal Ilmiah P2M STKIP Siliwangi*, *2*(1), 31–39. https://doi.org/10.22460/p2m.v2i1p31-39.161

Prijowuntato, S. W. (2016). *Evaluasi Pembelajaran*. Sanata Dharma University Press.

Rahmawati, D. O., Effendi, A., & Fatimah, A. T. (2013). Analisis Instrumen Tes Kemampuan Pemecahan Masalah Matematis Siswa pada Materi Trigonometri. *Prosiding Galuh Mathematics National Conference*, *3*(1), 28–35. https://jurnal.unigal.ac.id/GAMMA-NC/article/view/12948/6994

Rismen, S., Astuti, S., & Lovia, L. (2021). Analisis Kemampuan Pemahaman Konsep Matematis Siswa. *Lemma: Letters od Mathematics Education*, *7*(2), 123–134. https://doi.org/10.22202/jl.2021.v7i2.4911

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi

Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirāsāt: Jurnal Manajemen & Pendidikan Islam*, *2*(2), 192–213. https://doi.org/10.26594/dirasat.v2i2.879

Sudayana, R. (2015). *Statistika Penelitian Pendidikan*. Alfabeta.

Sugiyono. (2021). *Metode Penelitian Kuantitatif, Kualitatis, dan R&D*. Alfabeta.

Surapranata, S. (2009). *Analisis Validitas, Reliabilitas, dan Interpretasi Hasil Tes*. PT Remaja Rosdakarya.

Wulan, E. R., & Rusdiana, A. (2015). *Evaluasi Pembelajaran*. Pustaka Setia.