

## DEVELOPMENT OF A DIAGNOSTIC ASSESSMENT INSTRUMENT TO MEASURE THE MATHEMATICAL REASONING ABILITY OF JUNIOR HIGH SCHOOL STUDENTS

Siti Nurul Sajdah<sup>1</sup>, Kana Hidayati<sup>2</sup>

Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

[sitinurulsajdah@gmail.com](mailto:sitinurulsajdah@gmail.com), [kana@uny.ac.id](mailto:kana@uny.ac.id)

### Abstract

*This study aimed to produce a diagnostic assessment instrument for mathematical reasoning ability of junior high school students, describe the quality of diagnostic assessment instruments, and describe the profile of mathematical reasoning ability of junior high school students in Barru Regency based on the diagnostic assessment developed. This study used a development research approach referencing instrument development procedure from Edi Istiyono. Test and measurement subjects included 534 grade VIII students in Barru Regency. Data analysis included content validity through expert judgement analyzed using Aiken's  $V$  coefficient; construct validity through Confirmatory Factor Analysis (CFA); reliability estimation using Cronbach's Alpha coefficient; and item analysis based on item response theory with GPCM model. Results showed: (1) Construction of diagnostic assessment instruments for mathematical reasoning ability developed based on five aspects: compiling conjectures, providing reasons or evidence, drawing conclusions, examining validity of arguments, and finding patterns and making generalizations; (2) Instrument quality reviewed from content validity with  $V$  Aiken declared valid; construct validity tested through CFA, resulting in 19 items indicating loading factor  $> 0.35$ ; instrument reliability using Cronbach Alpha in high category; (3) Mathematical reasoning ability results were categorized as very high (18%), high (21%), low (41%), and very low (20%).*

*Keywords: Diagnostic Instruments, Mathematical Reasoning, IRT, Multiple Choice Two-Tier, Geometry.*

### Abstrak

*Penelitian ini bertujuan untuk menghasilkan instrumen asesmen diagnostik untuk kemampuan penalaran matematika siswa SMP, mendeskripsikan kualitas instrumen asesmen diagnostik, dan mendeskripsikan profil kemampuan penalaran matematika siswa SMP di Kabupaten Barru berdasarkan asesmen diagnostik yang dikembangkan. Penelitian ini menggunakan pendekatan penelitian pengembangan dengan mengacu pada prosedur pengembangan instrumen dari Edi Istiyono. Subjek pengujian dan pengukuran meliputi 534 siswa kelas VIII di Kabupaten Barru. Analisis data meliputi validitas isi melalui penilaian ahli yang dianalisis menggunakan koefisien Aiken  $V$ ; validitas konstruk melalui Analisis Faktor Konfirmatori (CFA); estimasi reliabilitas menggunakan koefisien Cronbach's Alpha; dan analisis item berdasarkan teori respons item dengan model GPCM. Hasil menunjukkan: (1) Konstruksi instrumen asesmen diagnostik untuk kemampuan penalaran matematika dikembangkan berdasarkan lima aspek: menyusun dugaan, memberikan alasan atau bukti, menarik kesimpulan, memeriksa validitas argumen, dan menemukan pola dan membuat generalisasi; (2) Kualitas instrumen ditinjau dari validitas isi dengan Aiken  $V$  dinyatakan valid; validitas konstruk diuji melalui CFA, menghasilkan 19 item yang menunjukkan faktor pemuatan  $> 0,35$ ; (2) Keandalan instrumen menggunakan Cronbach Alpha dalam kategori tinggi; (3) Hasil kemampuan penalaran matematika dikategorikan sebagai sangat tinggi (18%), tinggi (21%), rendah (41%), dan sangat rendah (20%).*

*Kata kunci: Instrumen Diagnostik, Penalaran Matematika, IRT, Pilihan Ganda Dua Tingkat, Geometri.*



© Author(s) 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

## INTRODUCTION

Mathematics can be a "vehicle" in developing critical, logical, analytical, and systematic abilities because mathematics is inseparable from the process of proof that requires reasoning and the use of strategies in problem-solving. In line with this, mathematical reasoning skills are needed to build students' mathematical abilities, thus facilitating them in understanding and solving problems. In this regard, mathematical reasoning plays a crucial role in students' capacity to understand mathematics and is emphasized more strongly in various international curricula. These abilities enable students to adapt to various challenges in competency development, as well as to acquire, manage, and utilize information effectively.<sup>1</sup>

Mathematical reasoning ability is an important aspect for students, data shows that Indonesian students still face significant challenges in this area, with a PISA score of 379 ranked 73rd and TIMSS results indicating that the abilities of eighth-grade junior high school students are in the low category in cognitive reasoning. Factors causing low student reasoning ability are learning that is still based on content mastery and memorization, as well as limited assessment instruments that can specifically measure students' reasoning abilities. Thus, an assessment is needed that can reveal the extent of students' understanding of mathematical concepts and mathematical reasoning.<sup>2</sup>

Based on Decree of the Minister of Education and Culture Number 719/P/2020, the government has established guidelines for curriculum implementation in educational institutions facing special circumstances. This regulation stipulates that diagnostic assessments must be conducted before the learning process begins. This diagnostic assessment aims to identify students' weaknesses and abilities that need improvement, so that learning plans can be tailored to the conditions and needs of each student. Theoretically, diagnostic assessments essentially function to identify students' strengths and weaknesses and provide input to teachers and students to make decisions related to improving the learning process. Thus, diagnostic assessments can help teachers identify differences in the abilities and learning needs of each student.<sup>3</sup>

The development of quality diagnostic instruments requires in-depth analysis to ensure the quality of the resulting instruments. *Item Response Theory* (IRT) is the most appropriate analysis approach because it has the advantages of parameter invariance, which allows item parameters to be independent of specific sample characteristics, adaptive measurement precision,

---

<sup>1</sup> Kuwat Triyana et al., "A Highly Sensitive Safrole Sensor Based on Polyvinyl Acetate (PVAc) Nanofiber-Coated QCM," *Scientific Reports* 9, no. 1 (2019): 15407.

<sup>2</sup> Rita Arfi Astuti Ningroom and Sri Yamtinah, "A Two-Tier Multiple-Choice Diagnostic Test to Find Student Misconceptions about the Change of Matter.," *Journal of Education and Learning (EduLearn)* 19, no. 2 (2025): 1144–56.

<sup>3</sup> Bashori Bashori, "Analisis Kebijakan Pendidikan Dalam Masa Darurat Penyebaran Covid-19 (Analisis SE Mendikbud No. 4 Tahun 2020)," *Jurnal Idarah: Pendidikan Dan Kependidikan* 5, no. 1 (2021).

which provides different levels of accuracy for each student's ability level, and a probabilistic model that provides a more comprehensive interpretation of the probability of students' success in answering questions based on their ability level. IRT allows the identification of the most informative items for each indicator of mathematical reasoning through item characteristic curves. Thus, teachers can obtain more comprehensive diagnostic information to design learning that suits individual students' needs.<sup>4</sup>

Based on the description above, this study focuses on developing a diagnostic assessment instrument specifically designed to measure junior high school students' mathematical reasoning abilities that is valid, reliable, and practical through IRT analysis. This study specifically selected geometry as the focus of instrument development. This selection is based on findings that reveal mathematical reasoning is closely related to spatial ability, where this ability is a fundamental component in understanding geometric concepts.<sup>5</sup>

## RESEARCH METHODS

This research is a development research that aims to produce a product in accordance with applicable standards. The product produced in this study is a diagnostic test instrument to measure students' mathematical reasoning abilities in mathematics.<sup>6</sup> The development model used in this study is the instrument development framework proposed by Istiyono which is generally divided into three stages: 1) test design which includes determining the test objectives, determining the competencies and materials to be tested, compiling the instrument matrix, preparing and compiling the test grid (*blueprint*), writing and assembling test items, preparing scoring rubrics, validity of test items, revisions to improve test items, assembling the instrument, 2) test trials which include determining the test subjects, conducting the test trials, scoring and analyzing the test items, revising test items that do not meet the desired test item parameter criteria, 3) measurement which includes assembling the test based on the test results, determining the measurement subjects, conducting the measurement, scoring and analyzing the measurement data, and interpreting the results. The development procedure refers to the Istiyono instrument development model.<sup>7</sup>

The test subjects in this study were eighth-grade junior high school students in Barru Regency. The research subjects were divided into two groups: the test subjects and the

---

<sup>4</sup> Tatan Zenal Mutakin et al., "Analisis Kemampuan Siswa Sekolah Dasar Dalam Menyelesaikan Soal Matematika Menggunakan Framework TIMSS 2019," *Diskusi Panel Nasional Pendidikan Matematika* 9 (2023).

<sup>5</sup> Marantika Lia Kristiyasari et al., "Analisis Validitas Isi Computerized Two-Tier Multiple Choice Untuk Mengukur Literasi Sains Siswa," 2017, 266–73.

<sup>6</sup> Rivo Panji Yudha, *Asesmen Unjuk Kerja Geometri* (Deepublish, 2020).

<sup>7</sup> Ronald K. Hambleton and Hariharan Swaminathan, "Estimation of Ability," in *Item Response Theory: Principles and Applications* (Springer, 1985).

measurement subjects, whose mathematical reasoning abilities were measured, as shown in Table 1 below.

**Table 1.** Research Subjects

No.	School Name	Number of Classes	Number of Student (Class VIII)	Jenis Sampel
1	UPTD SMP NEGERI 1 BARRU	9	261	Test Subjects
2	UPTD SMP NEGERI 2 BARRU	4	97	
3	UPTD SMP NEGERI 5 BARRU	6	176	Measurement Subject
<b>Amount</b>		19	<b>534</b>	

The test developed in this research is a diagnostic test in the form of *multiple choice two-tier*. Validation of the content of the research instrument was carried out by experts (*expert judgement*) involving seven experts who are a lecturer specializing in measurement, four lecturers specializing in mathematics, and two junior high school mathematics teachers. The results of the content validation assessment were analyzed using the Aiken V index. Proof of construct validity was carried out through confirmatory factor analysis or *Confirmatory Factor Analysis* (CFA). In addition, the reliability of the instrument was estimated using the Cronbach's Alpha index. The quality of the test items was analyzed using *Item Response Theory* (IRT) with model *Generalized Partial Kredit Model* (GPCM).<sup>8</sup>

## RESULT AND DISCUSSION

### Test Disign

#### 1. Determining Test Objectives

The instrument developed in this study aims to diagnose the mathematical reasoning abilities of Phase D grade VIII students in mathematics under the Independent Curriculum in Barru Regency, specifically geometry. This diagnostic instrument was

---

<sup>8</sup> Ulf Wiberg, "Future Perspectives for Sparsely Populated Areas in Sweden," *Rural Areas and Development* 2 (2004).

designed to reveal students' achievement levels and identify error patterns and difficulties experienced in completing the mathematical reasoning ability test.<sup>9</sup>

## 2. Determination of Competencies and Materials to be Tested

The competencies tested in this study were determined based on aspects of mathematical reasoning ability that had been analyzed through theoretical studies. Each aspect of mathematical reasoning ability served as a reference for measurable and observable indicators in the geometry domain. Aspects of mathematical reasoning ability include students' skills in formulating conjectures, presenting reasons or proofs, drawing conclusions, evaluating the validity of arguments, and identifying patterns and making generalizations.<sup>10</sup>

The material tested focuses on geometry, in accordance with the Learning Outcomes (CP) of the Independent Curriculum Phase D issued by the Ministry of Education and Culture of the Republic of Indonesia. The geometry tested includes parallel lines and angles, polygon angles, and congruence of geometric shapes, triangles, and quadrilaterals.<sup>11</sup>

## 3. Preparation of Item Distribution Matrix

The item distribution matrix was constructed to align the mathematical reasoning ability aspects with the geometry material being tested and to develop item numbers or codes. The resulting matrix can be seen in Table 9.

---

<sup>9</sup> Muhammad Salman Arrosyid et al., *Pengukuran & Evaluasi Pendidikan* (CV. Ruang Tentor, 2024).

<sup>10</sup> Nuryani Dwi Astuti et al., *Prinsip-Prinsip Pengukuran Dan Evaluasi Pendidikan: Disertai Dengan Contoh Kasus* (CV. Ruang Tentor, 2024).

<sup>11</sup> Ross J. Herbert et al., "Regional Variability of Aerosol Impacts on Clouds and Radiation in Global Kilometer-Scale Simulations," *Atmospheric Chemistry and Physics* 25, no. 14 (2025): 7789–814.

**Table 2.** Mathematical Reasoning Ability Diagnostic Test Question Matrix

Reasoning Aspect	Material				
	Parallel lines and angular lines	Polygonal angle	Congruence	Triangel	Rectangular
Develop conjectures or conjectures	1(B4)		1(B5)	1 (B8)	1 (B1)
Provide reasons or evidence	2 (B3, B7)			2 (B6, B9)	
Draw a conclusion	1 (B11)			3 (B10, B12, B13)	
Checking the validity of arguments			2 (B14, B15)	1(B20)	1(B16)
Finding patterns and making generalizations		1 (B18)		1 (B17)	1 (B19)

#### 4. Preparation of Test Grid

Developing a test outline is a key step in instrument development, serving as a reference for the question writing process, ensuring that all material and competencies are represented proportionally. The test outline components in this study include learning outcomes, mathematical reasoning ability, question indicators, cognitive levels, and item numbers.

## 5. Writing and Assembling Test Items

Based on the instrument matrix and grid that has been prepared, the next step is to write and assemble the test items, which is the implementation stage of the instrument grid that has been prepared.<sup>12</sup>

## 6. Test Item Content Validity

The results of the validity assessment by seven experts (*expert judgement*) analyzed using Aiken's V coefficient.

**Table 1.** Value Aiken's V

No. Butir	Nilai V	Ket.	No. Butir	Nilai V	Ket.
B1	1,00	Valid	B11	1,00	Valid
B2	1,00	Valid	B12	1,00	Valid
B3	1,00	Valid	B13	1,00	Valid
B4	1,00	Valid	B14	0,96	Valid
B5	1,00	Valid	B15	1,00	Valid
B6	1,00	Valid	B16	1,00	Valid
B7	1,00	Valid	B17	1,00	Valid
B8	1,00	Valid	B18	1,00	Valid
B9	1,00	Valid	B19	1,00	Valid
B10	0,96	Valid	B20	1,00	Valid

Based on the results of the content validation analysis using Aiken's V coefficient, the Aiken index value ranged from 0.96 to 1.00, indicating that the items developed and analyzed for content validity had a high level of validity. Therefore, the analysis results presented in Table 10 can be concluded that all 20 items developed were declared valid, and no items were dropped from the trial stage.

## Trial Phase

### 1. Construct Validity

The results of the instrument trial were analyzed for construct validity using *Confirmatory Factor Analysis* (CFA) on 358 students at SMP Negeri 1 Barru and SMP Negeri 2 Barru. The Kaiser-Meyer-Olkin (KMO) test result of 0.919 indicates sample adequacy for CFA analysis. The CFA results revealed that 19 of the 20 items had *factor loading* > 0.35 with *p-value* < 0.05. Item 1 was eliminated because it had *factor loading* of 0.250 < 0.35. Evaluation *Goodness of Fit* shows the model meets almost all the criteria with *Chi-square* = 273.61, GFI = 0.99, RMSEA = 0.04, CFI = 0.93, TLI = 1.00, AGFI = 0.99, and PNFI = 0.81.<sup>13</sup>

<sup>12</sup> Islamiani Safitri et al., *Teori Pengukuran Dan Evaluasi* (CV. Ruang Tentor, 2024).

<sup>13</sup> Safitri et al., *Teori Pengukuran Dan Evaluasi*.

## 2. Instrument Reliability Estimation

The estimation of the reliability of the instrument in this study uses the coefficient *Cronbach's Alpha* showed a value of  $\alpha=0.88$ , categorized as high. The minimum reliability coefficient value that must be achieved by the instrument is 0.7. Therefore, it can be concluded that the developed instrument has good consistency.

## 3. Item Response Theory

### a. Item Response Theory Assumption Test

#### 1) One-dimensional

The assumption of unidimensionality can be proven through factor analysis by paying attention to a number of indicators, such as values *Eigenvalue*, *Total Variance Explained*, And *Scree Plot*. This assumption is considered fulfilled if the test instrument shows the existence of one main factor that significantly dominates compared to other factors. Factor analysis to obtain the value *Eigenvalue* which can be observed in the section *Total Variance Explained* shows that the test instrument has one dominant factor of 5.893 with a cumulative percentage of 31.015%. The results of the analysis indicate the presence of one most dominant factor, and this finding is reinforced by the display *Scree Plot* which is shown below.

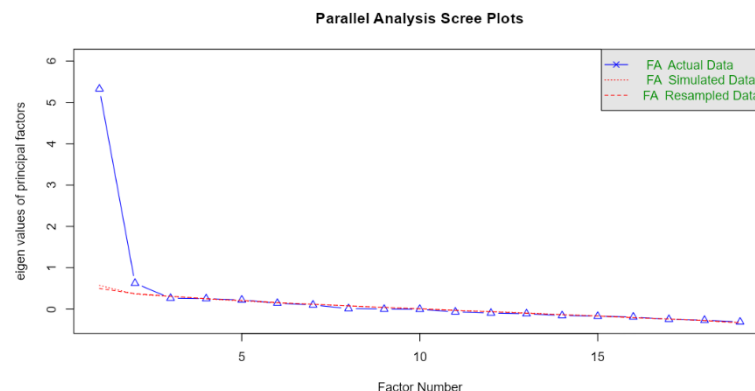


Figure 1. Unidimensional Scree Plot

Based on *Scree Plot* there was a sharp decline between factors 1 and 2, while the other factors began to plateau. Therefore, the diagnostic assessment instrument for measuring mathematical reasoning ability is unidimensional.

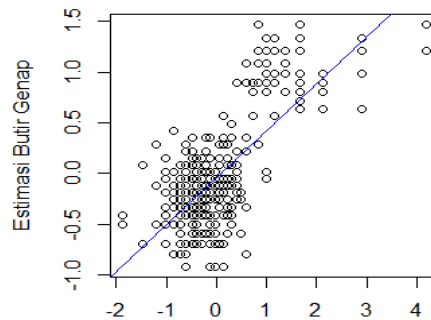
#### 2) Local Independence

The local independence assumption is considered met if a student's response to one item does not affect their response to another item. Furthermore, the fulfillment of the unidimensionality assumption also indicates that the local

independence assumption has been met. This is because these two concepts are equal and equivalent. In this context, because the unidimensionality assumption has been met, the local independence assumption is also automatically considered to have been met.<sup>14</sup>

### 3) Invariansi Parameter

The invariance of the difficulty level parameter describes the consistency of the parameter value which remains stable even when tested on two different groups of respondents.



**Figure 2.** Invariance of Ability Parameters

The invariance of the ability parameters in Figure 2 shows that the test takers' abilities did not change when working on only odd-numbered questions or when working on only even-numbered questions. Based on Figure 2, the invariance of the ability parameters has been met, so item analysis using the IRT approach can be conducted.<sup>15</sup>

#### b. Model Fit Test

Item response theory for polytomous data provides several models that can be applied, including: *Partial Credit Model* (PCM), *Generalized Partial Credit Model* (GPCM), and *Graded Response Model* (GRM). Thompson (2019) stated that if the data scoring system starts from a value of 0, then the most appropriate models to use are PCM and GPCM. This study uses a data scoring system starting from a value of 0, therefore the model used in the analysis is between PCM and GPCM. To determine the most appropriate model between PCM and GPCM, a comparative analysis was conducted based on several model suitability criteria, namely *Akaike Information*

<sup>14</sup> Eva Yanti Siregar et al., "Pengembangan Instrumen Penilaian Untuk Mengukur Kemampuan Penalaran Matematis Siswa Kelas VIII SMP N 2 Padangsidimpuan," *JURNAL MathEdu (Mathematic Education Journal)* 8, no. 2 (2025): 46–55.

<sup>15</sup> Yan Zhao and Baolu Zhao, "Oxidative Stress and the Pathogenesis of Alzheimer' s Disease," *Oxidative Medicine and Cellular Longevity* 2013, no. 1 (2013): 316523.

*Criterion (AIC), Bayesian Information Criterion (BIC)*, The model that produces the smallest AIC and BIC values is the most optimal model to use. The following are the results of the model fit analysis using AIC and BIC values.<sup>16</sup>

	AIC	SABIC	HQ	BIC
model_pcm	16348.39	16388.75	16436.36	16569.58
model_gpcm	16248.91	16302.72	16366.20	16543.83

**Figure 3.** Model Suitability Criteria

The model fit results above indicate that the model with the lowest AIC and BIC values is the GPCM model. Therefore, the item analysis in the next step will use the GPCM model.

### c. Item Parameter Estimation

Item parameter estimation can be performed after fulfilling the assumptions of item response theory and model fit testing. Based on the results of the model fit test, parameter estimation in this study was conducted using the GPCM model, with an emphasis on two main parameters: item difficulty (b) and item discrimination (a). A test item is considered good if the estimated difficulty level results are in the interval of -2.0 to 2.0 on the logit scale.

The results of the item difficulty analysis using the GPCM model show that there are three intersection points symbolized by b1, b2, and b3. The difficulty level of b1 ranges from -0.621 to 2.862. The difficulty level of b2 ranges from -1.947 to 1.361. The difficulty level of b3 is easier, at -1.286 to 1.470. Meanwhile, the value of *item location* ranges from -0.392 to 0.771, *item location* The highest is found in grains B13 and *item location* The lowest is found in item B19.<sup>17</sup>

The discrimination power of polytomous items ranges from 0.388 to 0.86. The item with the highest discrimination power is item B5, which measures the aspect of constructing conjectures. The indicator for item B5 is presenting two triangles, and students can determine whether the two shapes are similar using the angle information provided. Meanwhile, the polytomous item with the smallest discrimination power, 0.388, is item B13.<sup>18</sup>

---

<sup>16</sup> Novita Sari et al., *Konstruksi Instrumen Pendidikan* (CV. Ruang Tentor, 2025).

<sup>17</sup> Yudha, *Asesmen Unjuk Kerja Geometri*.

<sup>18</sup> Safitri et al., *Teori Pengukuran Dan Evaluasi*.

#### d. Grain Characteristic Curve

Grain characteristic curve or *Item Characteristic Curve* (ICC) is a graph that depicts the probability of respondents answering questions correctly based on their ability level. Parameter estimates can also be viewed through the ICC, which presents a functional relationship between respondents' abilities and the probability of success in answering test items, presented in graphical format.

The results of the analysis show that there are 19 questions that have *Item Characteristic Curve* (ICC) and covers five aspects of mathematical reasoning ability with a scoring range of 0 to 3. Examples of ICCs from several of these items are shown in Figure 4.

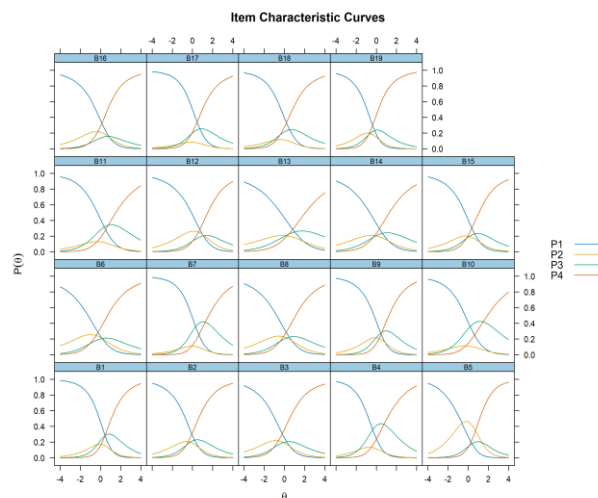
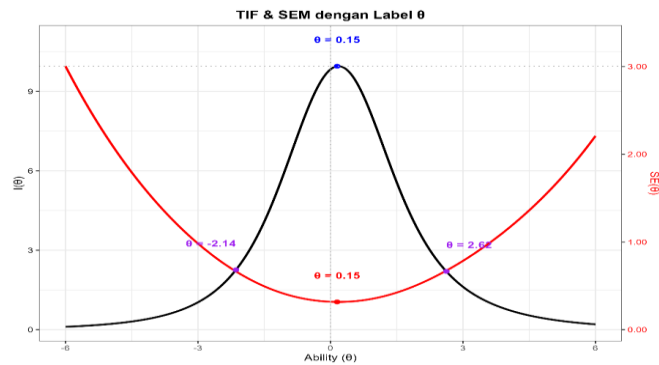


Figure 4. Grain Characteristic Curve

#### e. Item Information Function and *Standard Error of Measurement* (SEM)

The item information function is used to determine the extent to which each item contributes to measuring students' mathematical reasoning abilities. The resulting information can be used for the item selection process or to provide an overview of the item's strength in measuring latent abilities according to the stated measurement objectives. The results of the information function also make it possible to determine the score *Standard Error of Measurement* (SEM) or the level of error in the measurement process.



**Figure 5.** Standard Error of Measurement (SEM) curve

Based on Figure 5, the maximum value of the mathematical reasoning ability test information function is 10 at an ability of 0.15 and SEM of 0.317, meaning the test produces optimal information when used on students with an ability of -2.14. A good or reliable test instrument when the value *Total Information Function (TIF)*  $\geq 10$  According to Hambleton (Wiberg, 2004), the instrument used can be categorized as a good and reliable measuring tool in assessing students' mathematical reasoning abilities, as shown by the results of the analysis using the R program. In addition, there is an intersection of the information function curve and SEM intersecting at  $\theta = -2,14$  And  $2,62$  which means that this test as a whole is suitable for students with a range of abilities.  $\theta = -2,14$  And  $2,62$ . This range shows that the mathematical reasoning ability instrument is able to measure students' abilities over a fairly wide range.

### **Diagnostic Assessment Results to Measure the Mathematical Reasoning Ability of Junior High School Students in Barru Regency**

Students' mathematical reasoning abilities from measurement data were analyzed using the R program to produce ability estimates. The estimation of mathematical reasoning abilities in this study used the EAP approach with a range of values. *theta* for 176 students ranged from -1.76 to 2.01. The value *theta* then converted into *T-score* to determine a score on a scale of 0 to 100 according to school assessment standards. After that, categorization was carried out based on the Mardapi model. Details of the mathematical reasoning ability categories are shown in Table 4.

**Table 4.** Diagnostic Categories of Students' Mathematical Reasoning Abilities

<b>Mathematical Reasoning Ability Interval</b>	<b>Frequency</b>	<b>Percentage (%)</b>	<b>Category</b>
$X \geq 60,7$	35	18%	Very high
$60,7 > X \geq 51,3$	37	21%	High
$51,3 > X \geq 41,8$	72	41%	Low
$X < 41,8$	35	20%	Very Low
Amount	176	100%	

The implementation of the instrument with 19 questions on 176 class VIII students of SMP Negeri 5 Barru resulted in an estimate of ability with a range of  $\theta$ -1.76 to 2.01. The distribution of abilities shows that 18% of students are in the very high category, 21% in the high category, 41% in the low category, and 20% in the very low category. Based on this, 61% of students are in the very low to very low category, reflecting students' still weak mastery of mathematical reasoning skills.

Further analysis of each reasoning ability indicator revealed that students' primary weakness lay in the "finding patterns and making generalizations" indicator, with the lowest score of 21%. Students were also weak in the "drawing conclusions" indicator, with a score of 17%.

## Discussion

The final product developed in this study is a diagnostic assessment instrument to measure the mathematical reasoning abilities of eighth-grade junior high school students in geometry. This product is designed to assist teachers in diagnosing students' mathematical thinking weaknesses and strengths in more depth, allowing them to plan learning according to each student's learning needs.<sup>19</sup> This is in line with research by Triyono et al. which revealed that diagnostic tests *multiple choice two-tier* It can be said to be quite effective in identifying students' level of understanding of mathematical concepts. Furthermore, it can identify students'

<sup>19</sup> Hasan Basri and Abdur Rahman As' ari, "Investigating Critical Thinking Skill of Junior High School in Solving Mathematical Problem.," *International Journal of Instruction* 12, no. 3 (2019): 745–58.

misconceptions about material changes, improving knowledge, understanding, and critical thinking.<sup>20</sup>

The quality of the developed diagnostic assessment instrument shows quality in terms of validity, reliability, and item characteristics. Content validity, which was strengthened through the assessment of seven experts, showed that the instrument had met the criteria for validity based on the Aiken's  $V$  index value. This is reinforced by Yaghmaie's research which states that content validity is very important for developing valid instruments, because the instrument will accurately measure the content to be measured.<sup>21</sup>

Construct validity with *Confirmatory Factor Analysis* (CFA) which aims to confirm the suitability of the measurement model with empirical data. Apriandi et al. stated that there is a mismatch between the items and the measured variables if the value *factor loading* less than 0.3. Thus, item elimination was performed, resulting in 19 construct-valid items. The reliability of the diagnostic assessment instrument for measuring mathematical reasoning ability showed a high reliability value of 0.88, meeting the minimum criteria of 0.7. This provides confidence in the accuracy and consistency of the measurement results.<sup>22</sup>

Analysis of item quality using the GPCM model which focuses on the level of item difficulty ( $b$ ) and the differential power parameter ( $a$ ). Based on this analysis, the value *item location* The highest is found in item 13 with a value of ( $b = 0,771$ ), while the lowest value is found in item 19 with a value of ( $b = -0,392$ ). Thus, the instrument has a good level of difficulty, in accordance with the criteria of Hambleton & Swaminathan which stipulates the interval  $-2,0$  until  $2,0$ . In addition, the item discrimination parameter with the highest discrimination power is found in item B5 which measures the aspect of constructing conjectures. Item B5 has an indicator presented with two triangular shapes, students can determine whether the two shapes are similar by using the angle information provided. Meanwhile, the item with the smallest discrimination power of 0.388 is found in item B13.<sup>23</sup>

Diagnostic assessments enable teachers and schools to identify, monitor, and address students' mathematical reasoning difficulties, particularly in the reasoning aspect. Diagnostic assessments are assessments to determine students' strengths and weaknesses in mastering learning materials. By knowing the results of student ability diagnostics, teachers can design learning according to student needs. The results of diagnostic assessments of 176 eighth-grade

---

<sup>20</sup> Silvia Aras et al., "Kemampuan Kognitif Siswa Kelas Homogen Saat Mengerjakan Soal TIMSS Materi Bilangan," *Primatika: Jurnal Pendidikan Matematika* 13, no. 1 (2024): 55–62.

<sup>21</sup> Kristyasari et al., "Analisis Validitas Isi Computerized Two-Tier Multiple Choice Untuk Mengukur Literasi Sains Siswa."

<sup>22</sup> Zhao and Zhao, "Oxidative Stress and the Pathogenesis of Alzheimer's Disease."

<sup>23</sup> Siregar et al., "Pengembangan Instrumen Penilaian Untuk Mengukur Kemampuan Penalaran Matematis Siswa Kelas VIII SMP N 2 Padangsidempuan."

junior high school students in Barru Regency indicate that students' mathematical reasoning abilities are still quite diverse, with ability categories tending to be in the low to very low categories. Based on this, as many as 61% of students are in the very low to very low category, which reflects students' still weak mastery of mathematical reasoning abilities.

Further findings on each indicator of mathematical reasoning ability revealed that students' primary weakness lies in the indicator "finding patterns and making generalizations," with the lowest score of 21%. Students were also weak in the indicator "drawing conclusions," which scored 17%. This indicates that students have difficulty identifying patterns, developing generalizations from specific patterns, and drawing valid conclusions based on the information or arguments provided.

These findings are particularly concerning, given that reasoning is the foundation of mathematics learning. If students' reasoning skills are not developed, mathematics will become a mere matter of following a series of procedures and imitating examples without understanding their meaning.<sup>24</sup>

This situation is reinforced by the findings of Syahlan et al., who revealed that currently, students tend to work on problems given based on examples given by the teacher, but when asked to develop ideas to solve problems, students experience difficulties. This indicates low student reasoning abilities and can lead to decreased mathematics learning outcomes. Therefore, the diagnostic assessment instrument for measuring mathematical reasoning abilities has met the instrument development stages and is declared valid and reliable, thus evaluating students' mathematical reasoning abilities in mathematics learning.<sup>25</sup>

## CONCLUSION

Based on the results of research and product studies, the following conclusions can be drawn.

1. Instrument construction *multiple choice two-tier*. The questionnaire consists of 20 questions to measure the mathematical reasoning abilities of eighth-grade junior high school students in geometry. The instrument addresses five aspects of mathematical reasoning, with the first level being multiple-choice questions and the second level being logical reasoning, which can reveal students' thought processes.
2. The instrument quality was declared good based on content validity with Aiken's V ranging from 0.96 to 1.00, construct validity CFA yielded 19 valid items out of 20,

---

<sup>24</sup> Nurul Hasanah, "Pengembangan Lembar Kerja Peserta Didik Berbasis Karakter," *Jurnal Pendidikan Dan Pembelajaran Terpadu* 1, no. 1 (2019): 24–30.

<sup>25</sup> Nusrotus Sa'idah and Hayu Dian Yulistianti, "Pengembangan Modul Praktikum Berbasis Analisis Data Pada Mata Kuliah Statistik Dalam Meningkatkan Belajar Mandiri Mahasiswa," *Kreano, Jurnal Matematika Kreatif-Inovatif* 9, no. 2 (2018): 198–203.

Cronbach's Alpha reliability was 0.88, and IRT-GPCM analysis showed adequate difficulty and discrimination parameters. The instrument is suitable for use as a diagnostic measurement tool.

3. Diagnostic results from 176 students showed that the majority were in the low to very low category in mathematical reasoning ability. The lowest indicators were finding patterns, making generalizations, and drawing conclusions. The instrument can identify specific student weaknesses and serve as a basis for lesson planning.

## BIBLIOGRAPHY

- Aras, Silvia, Rusdiana Rusdiana, and Nanda Arista Rizki. "Kemampuan Kognitif Siswa Kelas Homogen Saat Mengerjakan Soal TIMSS Materi Bilangan." *Primatika: Jurnal Pendidikan Matematika* 13, no. 1 (2024): 55–62.
- Arrosyid, Muhammad Salman, Andi A. Manggaberani, Arief Nurrahman, et al. *Pengukuran & Evaluasi Pendidikan*. CV. Ruang Tentor, 2024.
- Astuti, Nuryani Dwi, Amran Hapsan, Mutmainna Herianto, et al. *Prinsip-Prinsip Pengukuran Dan Evaluasi Pendidikan: Disertai Dengan Contoh Kasus*. CV. Ruang Tentor, 2024.
- Bashori, Bashori. "Analisis Kebijakan Pendidikan Dalam Masa Darurat Penyebaran Covid-19 (Analisis SE Mendikbud No. 4 Tahun 2020)." *Jurnal Idarah: Pendidikan Dan Kependidikan* 5, no. 1 (2021).
- Basri, Hasan, and Abdur Rahman As' ari. "Investigating Critical Thinking Skill of Junior High School in Solving Mathematical Problem." *International Journal of Instruction* 12, no. 3 (2019): 745–58.
- Hambleton, Ronald K., and Hariharan Swaminathan. "Estimation of Ability." In *Item Response Theory: Principles and Applications*. Springer, 1985.
- Hasanah, Nurul. "Pengembangan Lembar Kerja Peserta Didik Berbasis Karakter." *Jurnal Pendidikan Dan Pembelajaran Terpadu* 1, no. 1 (2019): 24–30.
- Herbert, Ross J., Andrew IL Williams, Philipp Weiss, et al. "Regional Variability of Aerosol Impacts on Clouds and Radiation in Global Kilometer-Scale Simulations." *Atmospheric Chemistry and Physics* 25, no. 14 (2025): 7789–814.
- Mutakin, Tatan Zenal, Burhanuddin Tola, and Bahrul Hayat. "Analisis Kemampuan Siswa Sekolah Dasar Dalam Menyelesaikan Soal Matematika Menggunakan Framwork TIMSS 2019." *Diskusi Panel Nasional Pendidikan Matematika* 9 (2023).
- Ningroom, Rita Arfi Astuti, and Sri Yamtinah. "A Two-Tier Multiple-Choice Diagnostic Test to Find Student Misconceptions about the Change of Matter." *Journal of Education and Learning (EduLearn)* 19, no. 2 (2025): 1144–56.
- Safitri, Islamiani, Dewi Lestarani, Rahmah Dwi Nor Wita Imtikhanah, et al. *Teori Pengukuran Dan Evaluasi*. CV. Ruang Tentor, 2024.
- Sa'idah, Nusrotus, and Hayu Dian Yulistianti. "Pengembangan Modul Praktikum Berbasis Analisis Data Pada Mata Kuliah Statistik Dalam Meningkatkan Belajar Mandiri Mahasiswa." *Kreano, Jurnal Matematika Kreatif-Inovatif* 9, no. 2 (2018): 198–203.
- Sari, Novita, Andi Abdurrahman Manggaberani, Andi Jusmiana, et al. *Konstruksi Instrumen Pendidikan*. CV. Ruang Tentor, 2025.

Siti Nurul Sajdah, Kana Hidayati: Development of a Diagnostic Assessment Instrument to Measure the Mathematical Reasoning Ability of Junior High School Students

Siregar, Eva Yanti, Rahmatika Elindra, and Vivin Mei Sarah Sikumbang. "Pengembangan Instrumen Penilaian Untuk Mengukur Kemampuan Penalaran Matematis Siswa Kelas VIII SMP N 2 Padangsidempuan." *JURNAL MathEdu (Mathematic Education Journal)* 8, no. 2 (2025): 46–55.

Triyana, Kuwat, Aditya Rianjanu, Doni Bowo Nugroho, et al. "A Highly Sensitive Safrrole Sensor Based on Polyvinyl Acetate (PVAc) Nanofiber-Coated QCM." *Scientific Reports* 9, no. 1 (2019): 15407.

Wiberg, Ulf. "Future Perspectives for Sparsely Populated Areas in Sweden." *Rural Areas and Development* 2 (2004).

Yudha, Rivo Panji. *Asesmen Unjuk Kerja Geometri*. Deepublish, 2020.

Zhao, Yan, and Baolu Zhao. "Oxidative Stress and the Pathogenesis of Alzheimer' s Disease." *Oxidative Medicine and Cellular Longevity* 2013, no. 1 (2013): 316523.