Journal of Mathematics Instruction, Social Research and Opinion
Vol. 4, No. 4, December 2025, pp. 1453 – 1464, https://doi.org/10.58421/misro.v4i4.873
ISSN 2962-7842

Psychometric Evaluation of a Metaphorical Thinking Instrument in Mathematics Learning: Graded Response Model

Oryza Vini Faradila[1], Ilham Falani[2], Mujahidawati[3]
[1,2,3]Universitas Jambi, Jambi, Indonesia

Article Info

Article history:

Received 2025-12-11
Revised 2025-12-23
Accepted 2025-12-24

Keywords:

Graded Response Model
Item Response Theory
Metaphorical thinking
Validity and reliability

ABSTRACT

This study aims to develop and psychometrically validate an assessment instrument designed to measure students' metaphorical thinking ability in mathematics learning, with a specific focus on the Pythagorean Theorem. The instrument was developed using a Research and Development (R&D) framework based on the Oriondo and Antonio model, encompassing test design, empirical tryout, and measurement stages. Six polytomous essay items were constructed according to six indicators of metaphorical thinking: connect, relate, explore, analyze, transform, and experience. Content validity was established through expert judgment using Aiken's V coefficient, with all items exceeding the minimum validity threshold, indicating strong agreement among experts. Empirical validation was conducted using Item Response Theory (IRT) with the Graded Response Model (GRM), selected for its suitability in analyzing ordered polytomous response data. The results demonstrate that the instrument satisfies the unidimensionality assumption, exhibits strong item discrimination parameters, and shows good model fit across all items. Analysis of the Test Information Function indicates high measurement precision within the ability range of θ −1 to +1, confirming strong local reliability. These findings indicate that the developed instrument is valid, reliable, and capable of providing accurate diagnostic information regarding students' ability to construct mathematical meaning through metaphors. The study contributes methodologically by demonstrating the applicability of GRM-based IRT analysis for essay-type instruments and substantively by supporting the assessment of higher-order cognitive processes in mathematics learning.

Corresponding Author:

Ilham Falani
Master of Mathematics Education, Faculty of Teacher Training and Education, Universitas Jambi
Email: ilhamfalani@unja.ac.id

1. INTRODUCTION

Assessment in mathematics education plays a strategic role not only in measuring learning outcomes but also in supporting instructional decision-making, diagnosing students' cognitive difficulties, and fostering meaningful learning processes [1].

Contemporary assessment paradigms emphasize the need to capture complex cognitive abilities beyond procedural competence, particularly those related to conceptual understanding and meaning construction [2]. One such ability that has received increasing theoretical attention but remains under-assessed in classroom practice is metaphorical thinking [3], [4], [5].

Metaphorical thinking functions as a cognitive bridge that helps students connect mathematical ideas with real-world representations, thereby facilitating the internalization of concepts [6], [7], [8]. Previous studies have shown that the use of metaphors in learning can enhance students' conceptual understanding, creativity, and mathematical reasoning [9], [10]. Despite the growing theoretical recognition of metaphorical thinking as a crucial cognitive ability for constructing mathematical meaning, its assessment in mathematics education remains limited and underdeveloped. Existing classroom assessments and research studies predominantly focus on procedural performance, general problem-solving skills, or broad higher-order thinking abilities, while students' capacity to interpret and connect mathematical concepts through metaphors is rarely evaluated systematically [11]. Moreover, studies that attempt to assess metaphorical thinking often rely on classical test theory or descriptive approaches, which provide limited information regarding item functioning and measurement precision. From a psychometric standpoint, the application of Item Response Theory, particularly the Graded Response Model suited for ordered polytomous responses such as essay-based scoring, has received minimal attention in the context of metaphorical thinking assessment. This condition indicates a clear substantive and methodological gap in mathematics education research, highlighting the need for a validated, IRT-based instrument capable of providing precise and diagnostic measurement of students' metaphorical thinking ability [12], [13].

Findings from preliminary studies and interviews with mathematics teachers indicate that there is no standardized instrument specifically designed to assess metaphorical thinking in the context of mathematics learning. Teachers acknowledge that metaphorical approaches effectively support students in understanding abstract concepts, but limited theoretical knowledge and the absence of valid and reliable instruments hinder the systematic implementation of such assessments. This condition reveals a gap between classroom practice and the availability of appropriate measurement tools [14], [15], [16].

From a methodological perspective, Item Response Theory (IRT) offers a robust and modern approach to developing assessment instruments [17]. IRT generates more accurate item parameters and ability estimates that are invariant to sample characteristics and test forms, making it superior to classical approaches [18], [19], [20]. Nevertheless, previous studies employing IRT have primarily focused on general cognitive abilities and have not been directed toward assessing metaphorical thinking [21], [22]. Therefore, designing an IRT-based instrument for assessing metaphorical thinking represents an important opportunity to fill a gap in mathematics assessment research [23], [24].

This study aims to design and validate an assessment instrument for metaphorical thinking using the Graded Response Model (GRM) within the IRT framework. The instrument was constructed based on six key indicators of metaphorical thinking connect, relate, explore, analyze, transform, and experience, and was then empirically tested to obtain

evidence of validity, reliability, and item quality [4], [11]. The study offers innovative contributions by (i) providing an assessment instrument specifically aimed at measuring students' ability to construct mathematical meaning through metaphors, and (ii) applying modern measurement theory to produce an assessment that is more objective, diagnostic, and targeted.

Thus, this research not only addresses the need for an assessment instrument aligned with the evolving practices of mathematics education but also enriches the literature on metaphorical thinking assessment, which has received limited attention in mathematics education research.

2. METHOD

This study employed a research and development (R&D) design based on the model proposed by Oriondo and Antonio, as the framework is specifically oriented toward the systematic construction and validation of educational assessment instruments. The model emphasizes structured stages of test planning, item development, expert validation, pilot testing, and empirical analysis, ensuring that the resulting instrument fulfills both content validity and psychometric quality requirements [25]. Given that the primary focus of this study is the development of a valid and reliable instrument to assess students' metaphorical thinking ability in mathematics, rather than instructional intervention, the Oriondo and Antonio model is considered methodologically appropriate. Furthermore, data analysis was conducted using the Graded Response Model (GRM) within the Item Response Theory framework, as the instrument consisted of essay-type items scored using ordered polytomous categories based on an analytic rubric. GRM is particularly suitable for modeling ordered response categories and estimating category boundary parameters, allowing for a more precise representation of students' varying levels of metaphorical thinking ability. Compared to the Partial Credit Model and Rasch-based models, which assume equal item discrimination across items, GRM provides greater flexibility by allowing discrimination parameters to vary, making it more capable of capturing differences in cognitive demand and interpretative complexity inherent in essay-based assessments. Consequently, the application of GRM enables more accurate parameter estimation and richer diagnostic information for measuring students' metaphorical thinking ability [26].

2.1 Research Design

This study is a Research and Development (R&D) study that produces a product in the form of an assessment instrument for measuring metaphorical thinking ability based on Item Response Theory (IRT). A quantitative approach was employed in the item analysis stage to estimate item parameters, examine item fit, and calculate the reliability of the instrument using the Graded Response Model (GRM) [19].

2.2 Research Subjects and Location

The study was conducted with eighth-grade students at SMP Negeri 7 Muaro Jambi, located in the Jambi Luar Kota District, Muaro Jambi Regency, during the first semester of

the 2025/2026 academic year. The sample was selected purposively based on schools that were willing to collaborate and were relevant to the needs of the study.

2.3 Research Procedure and

The research procedure was carried out through three main stages as outlined by Oriondo and Antonio [26]. The research followed three main stages. The first stage focused on designing the instrument, beginning with defining its purpose, assessing students' metaphorical thinking in learning the Pythagorean theorem. Competencies were based on six indicators, which guided the development of the instrument blueprint and polytomous essay items aligned with the GRM model. Expert validation was then conducted, followed by revisions based on the feedback received.

The second stage involved trying out the instrument to collect empirical data. The test was administered according to school procedures, responses were scored using an analytic rubric, and item analysis using IRT-GRM was performed to examine step difficulty, item fit, category functioning, and ability distribution. These results informed decisions on whether items should be retained, revised, or removed.

The final stage included measurement and interpretation. The refined instrument was assembled into its final form and administered again to obtain accurate ability estimates. Reliability was assessed using person and item reliability indices within the IRT framework. The results were then interpreted through ability mapping, score category analysis, and continuous evaluation to ensure the instrument’s long-term effectiveness.

2.4 Data Collection Technique

The data collection technique in this research was carried out through several primary sources. Data were obtained from students’ responses to the test instrument administered during the trial phase, which served as the basis for evaluating item quality and assessing student abilities. In addition, expert feedback from the validation process was used to evaluate the construct appropriateness, linguistic clarity, and content accuracy of the instrument. Scoring was conducted using scoring sheets developed according to an analytic rubric, ensuring that each student’s response was assessed systematically based on predetermined indicators. Documentation gathered during the trial process was also utilized to provide contextual insights, ensure procedural fidelity, and strengthen the interpretation of the collected data. Altogether, these data collection techniques complemented one another to produce comprehensive and accurate information.

2.5 Data Analysis Technique

Data analysis in this study was conducted using two main approaches: qualitative and quantitative. Qualitative analysis was applied during the expert validation process to assess the construct, content, language, and feasibility of the instrument. This validation was carried out using Aiken’s V index to measure the degree of agreement among experts for each item, with higher values indicating stronger content validity. Feedback from the experts was then used as the basis for revising the instrument before it was trialed with students.

Meanwhile, quantitative analysis was conducted using the Item Response Theory (IRT) approach with the Graded Response Model (GRM) developed by Samejima. The GRM was chosen because it is suitable for ordered polytomous responses, such as rubric scores used in the metaphorical thinking instrument. This approach allows researchers to obtain detailed information about the characteristics of each score category within an item and to model the probability of students selecting a particular category based on their latent ability (theta). By combining qualitative analysis based on Aiken's V and quantitative GRM analysis, the instrument used is not only content-valid but also provides highly precise measurement.

3. RESULTS AND DISCUSSION

3.1. Results

The result of the instrument design research conducted was a metaphorical thinking test instrument for mathematics learning on the topic of the Pythagorean theorem. The implementation of the research results on instrument design followed the stages of the Oriondo & Antonio model [26], as described below.

Test Design

In the test design stage, the first step taken by the researcher was to determine the test objectives. The next step involved defining the competencies to be assessed. The competencies used were the Learning Outcomes (CP) and Learning Objectives (TP) for eighth-grade mathematics in junior high school during the odd semester of the 2025/2026 academic year.

Table 1. Learning Outcomes and Learning Objectives

Learning Outcomes (CP)	Learning Objectives (TP)
At the end of phase D, students can explain the concept of the Pythagoras theorem and solve real-world problems involving the relationships between the sides of a right-angled triangle.	Students can solve real-world problems using the Pythagorean theorem.

Next, in the material determination stage, the researcher focused the instrument design on the topic of the Pythagorean theorem. This topic was chosen because it is one of the subjects studied in mathematics learning. Its characteristics provide opportunities for students to demonstrate metaphorical thinking skills, which is the ability to connect abstract concepts with more concrete representations or experiences. Then, the instrument blueprint was developed by referring to the metaphorical thinking ability indicators proposed by Siller, which include six main components: Connect, Relate, Explore, Analyze, Transform, and Experience. Next, the assessment instrument in this study consisted of six essay questions grouped into a single set. Each question was developed based on the previously prepared blueprint to measure metaphorical thinking skills. For test validation, the researcher provided six mathematics essay questions to experts in the field of mathematics. Validation was conducted using Aiken's method. The V_table value of 0.8 indicated that all items exceeded the minimum threshold, allowing the conclusion that all questions had a good level of validity. Item revision was carried out by referring to the comments and suggestions

provided by the validators. Each piece of feedback was carefully reviewed to ensure that the test items accurately reflected the concept of metaphorical thinking in mathematics learning, particularly on the topic of the Pythagorean theorem.

Test Trial Results

The trial subjects in this study were eighth-grade students at SMP Negeri 7 Muaro Jambi. Eighth grade was selected because the material studied aligns with the competencies assessed by the instrument. The test trial was conducted from November 10 to November 18, 2025, at SMP Negeri 7 Muaro Jambi, involving eight classes, labeled A to H. The activity was carried out directly by the researcher under the supervision of subject teachers to ensure the orderly conduct of the process and adherence to research procedures. The purpose of this trial was to obtain empirical data regarding item quality, readability, and the consistency of the instrument before it was used in the main data collection phase.

IRT Assumption Test Results

Before further data analysis was conducted using PARSCALE 4.1, it was necessary to test the prerequisite assumptions of IRT. The unidimensionality assumption can be tested using factor analysis with the aid of the Statistical Package for the Social Science (SPSS) version 26. The unidimensionality test was performed on the data prior to estimating the test participants' ability parameters. The data tested were polytomous data intended for the GRM model.

Table 2. Total Variance Explained

	Initial Eigenvalues			*Extraction Sums of Squared Loadings*		
Component	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	3,494	58,238	58,238	3,494	58,238	58,238
2	0,769	12,815	71,053			
3	0,661	11,018	82,071			
4	0,492	8,203	90,274			
5	0,362	6,033	96,307			
6	0,222	3,693	100,000			

Extraction Method: Principal Component Analysis.

Table 2 above shows the results of the unidimensionality test using Principal Component Analysis (PCA). It can be seen that the first component has an eigenvalue (Total) of 3.494, which accounts for 58.238% of the total variance explained by the data. This value is significantly higher than the other components, each of which has an eigenvalue of less than 1. The following image shows the scree plot graph of the data.

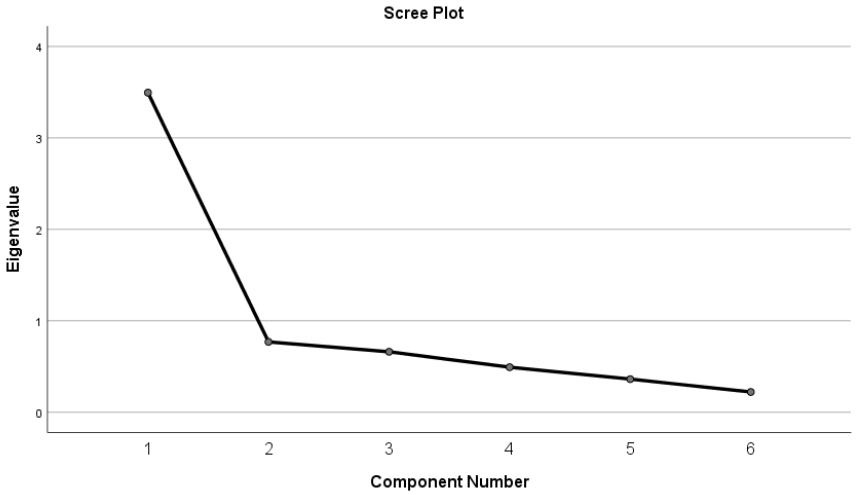

Figure 1. Scree Plot Data GRM

The scree plot graph was used to examine the dimensionality of the instrument using Principal Component Analysis (PCA).

Validity of the Metaphorical Thinking Assessment Instrument

Table 3. Correlations, GRM Parameters, and Item Difficulty

Item	Pearson Correlation	Polyserial Correlation	Slope (a)	SE (a)	Difficulty (b)	SE (b)
1	0,378	0,422	2,266	1,459	-1,094	0,046
2	0,565	0,641	1,835	0,463	-1,074	0,050
3	0,867	0,932	2,564	1,946	-0.335	0,023
4	0,700	0,865	1,724	0,422	-1,500	0,048
5	0,864	0,926	2,461	2,341	-0,058	0,025
6	0,830	0,889	1,938	0,641	0,135	0,026

The internal validity of the instrument was analyzed through the correlation between each item score and the total score, reflected by Pearson correlation and polyserial correlation values. The analysis results showed that all items fell within a strong psychometric correlation range, with Pearson correlations between 0.378 and 0.867 and polyserial correlations between 0.422 and 0.932. These ranges indicate internal consistency and a substantial relationship between participants' responses and the metaphorical thinking construct being measured. Based on methodological criteria, correlation values above 0.30 meet the internal validity requirement, so all items can be considered empirically valid.

Table 4. Item Fit

Item	Chi-Square	df	Prob	Keterangan
1	7,84707	12	0,864	FIT
2	5,39957	12	0,943	FIT
3	8,86816	12	0,725	FIT
4	5,11951	11	0,928	FIT
5	8,60678	13	0,812	FIT
6	7,42131	14	0,917	FIT

The item fit analysis results indicated that all items in the instrument met the fit requirements for the IRT model used. Item fit was evaluated based on the chi-square value, degrees of freedom (df), and probability (p-value). In the context of IRT, an item is considered fit if the p-value exceeds the significance threshold of 0.05, indicating no significant deviation between the empirical response patterns and those predicted by the model. Thus, the model is able to accurately represent the characteristics of the items.

Item Characteristic Curve (ICC)

The Item Characteristic Curve (ICC) represents the relationship between respondents' latent ability and the probability of providing a correct response to each item.

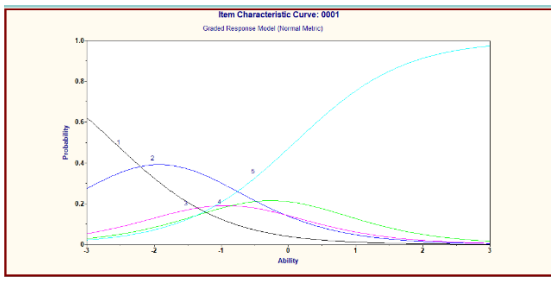

Figure 2. ICC item 1

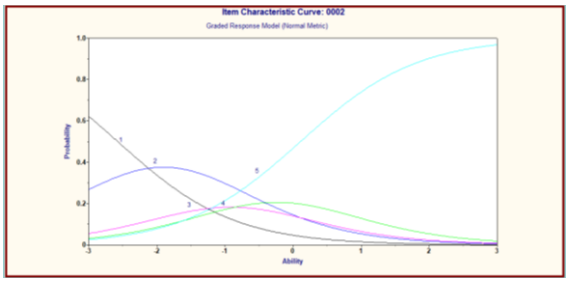

Figure 3. ICC item 2

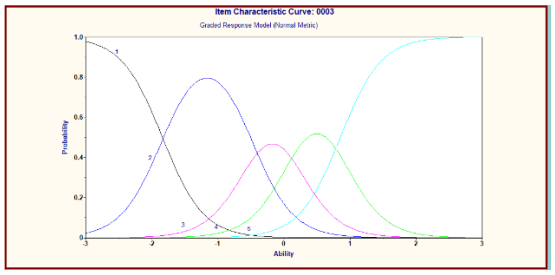

Figure 4. ICC item 3

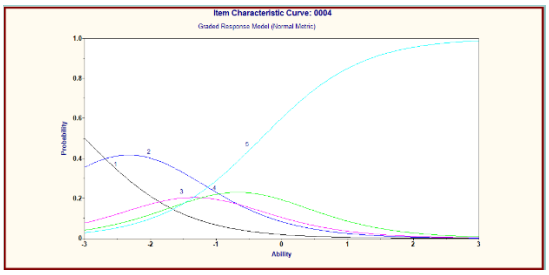

Figure 5. ICC item 4

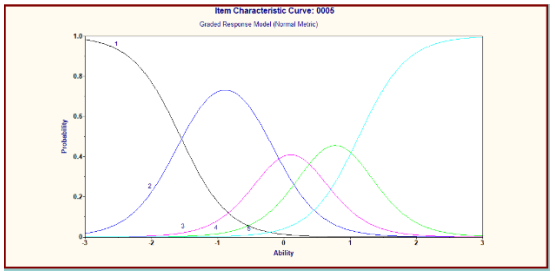

Figure 6. ICC item 5

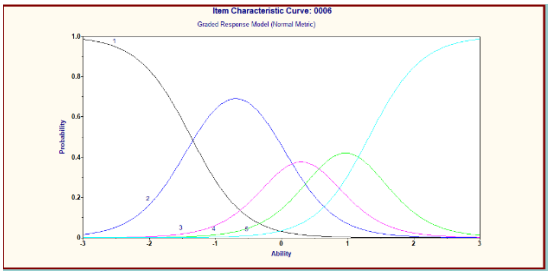

Figure 7. ICC item 6

Test Information Function

The Item Response Theory (IRT) analysis results, based on the Test Information Function (TIF) and Standard Error of Measurement (SEM) outputs, showed that the instrument has excellent measurement properties within a certain ability range.

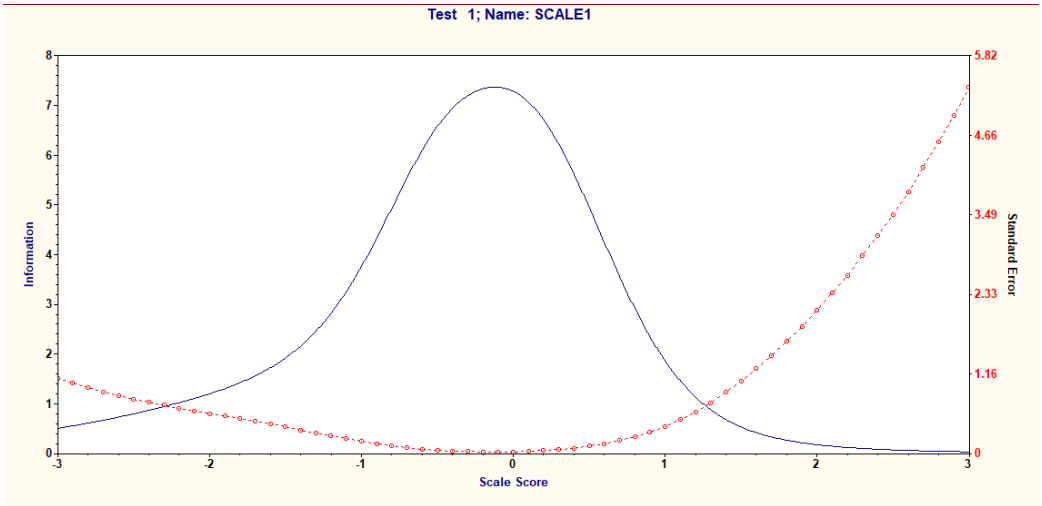

Figure 8. Test Information Function (TIF)

The TIF curve, shown as a solid blue line, indicates that the maximum information value reaches approximately 7. This TIF–SEM pattern demonstrates that the instrument has strong reliability within the targeted ability range. A high information peak and low SEM indicate that the instrument can accurately differentiate participants within a population of average ability. Therefore, the local reliability of the instrument in the θ range between −1 and +1 can be categorized as very good. The test information function indicates that the developed instrument provides optimal measurement precision at low to moderate levels of students' metaphorical thinking ability. This finding suggests that the instrument is particularly effective for diagnostic purposes, as it is able to capture meaningful differences in students' ability to construct mathematical meaning through metaphorical reasoning. Furthermore, the variation in item discrimination parameters reflects differences in items' sensitivity to distinguish students across ability levels, supporting the appropriateness of applying a flexible IRT-based model for essay-type assessments.

3.2. Discussion

The instrument design process in this study demonstrates that the metaphorical thinking assessment requires the application of modern instrument design principles oriented toward integrating theoretical foundations, curriculum needs analysis, and methodological accuracy. Instrument development cannot be carried out solely through technical item construction but must consider the alignment between measurement objectives, the structure of learning competencies, and the psychometric characteristics of the instrument to be analyzed. The findings of this study provide empirical support for the theoretical assumption that metaphorical thinking constitutes a coherent latent construct that can be meaningfully assessed through students' written mathematical responses. The ability of the instrument to differentiate students' levels of metaphorical thinking aligns with the view that metaphor use plays a central role in facilitating conceptual understanding and meaning construction in mathematics learning.

The findings of this study reinforce the theoretical perspective that metaphorical thinking is a fundamental cognitive mechanism in mathematical learning, enabling students to construct meaning by mapping abstract mathematical concepts onto more familiar or

concrete experiences. The ability of the instrument to distinguish varying levels of metaphorical thinking empirically supports the view that this ability represents a coherent latent construct rather than a fragmented set of skills. This result aligns with cognitive and educational theories that emphasize the role of metaphor in facilitating conceptual understanding, abstraction, and sense-making processes in mathematics [4], [19].

The validity of the instrument in this study was analyzed using two complementary approaches: content validity based on expert evaluation and empirical construct validity based on Item Response Theory (IRT) analysis [19]. Expert assessment indicated that all instrument items had Aiken's V values above 0.80, which falls within the valid category [27]. This confirms that experts agreed on the alignment between the metaphorical thinking indicators and item content, the clarity of sentence structures, the appropriateness of metaphor usage, and the relevance of the items to the Pythagorean Theorem material.

The construct validity of the instrument was further supported by empirical analysis using the Graded Response Model (GRM) within the IRT framework [19]. Initial analysis showed that the instrument met the unidimensionality assumption, with a single primary component explaining 58.238% of the total variance. This finding indicates that all items measure the same latent construct, namely metaphorical thinking. Construct validity was further reinforced by additional empirical evidence. Pearson and polyserial correlations ranged from 0.378 to 0.932. Moreover, the item fit results indicated that no items were misfitting, as all p-values exceeded 0.05. The IRT analysis using the GRM model showed that the instrument has a strong construct structure, stable item performance, and optimal model fit. From a methodological perspective, the successful application of the Graded Response Model demonstrates its suitability for analyzing essay-based assessments involving ordered polytomous scoring. Compared to Rasch-based and Partial Credit models, which impose equal discrimination assumptions, GRM offers greater flexibility in modeling item characteristics, making it more capable of capturing the complexity of higher-order cognitive abilities such as metaphorical thinking.

In terms of pedagogical implications, the availability of a validated instrument for assessing metaphorical thinking enables teachers to move beyond evaluating procedural accuracy toward diagnosing students' conceptual meaning-making processes. By identifying students' strengths and difficulties in using metaphors to interpret mathematical ideas, educators can design more targeted instructional interventions that support deeper understanding and conceptual development in mathematics learning.

Reliability analysis based on the IRT approach indicated that the metaphorical thinking assessment instrument developed in this study possesses high measurement precision and consistency within the most pedagogically relevant ability range. The instrument is suitable for both further research and practical assessment in mathematics learning, as it can provide accurate estimates of students' metaphorical abilities and support data-driven decision-making in the context of learning evaluation.

4. CONCLUSION

Based on the analysis using Item Response Theory (IRT) with the Graded Response Model (GRM), the metaphorical thinking assessment instrument in mathematics learning

developed in this study has been proven to be valid, reliable, and psychometrically sound. These findings indicate that the instrument can consistently measure students' ability to understand and apply mathematical metaphors, particularly in the context of the Pythagorean Theorem. Furthermore, the instrument can not only be used to assess students' skills in constructing, connecting, and interpreting mathematical concepts creatively but also has the potential to serve as a foundation for developing innovative assessments in mathematics education. Therefore, this instrument makes a significant contribution to strengthening the quality of learning evaluation and promotes the development of more contextual and reflective learning strategies for students. This study contributes to mathematics education research by providing an empirically validated assessment instrument for measuring students' metaphorical thinking ability using an Item Response Theory–based approach. The instrument offers practical value for diagnostic assessment by enabling teachers and researchers to identify students' conceptual meaning-making processes beyond procedural performance. Nevertheless, the findings are limited to a specific mathematical topic and sample context, suggesting that future research should further examine the applicability of the instrument across different content domains and educational settings.

ACKNOWLEDGEMENTS

The authors would like to express their sincere appreciation and gratitude to Dr. Ilham Falani, S.Pd., M.Si., and Dr. Dra. Mujahidawati, M.Si. for their guidance, direction, and valuable input throughout the preparation of this article. The authors also extend their thanks to the Master's Program in Mathematics Education at the University of Jambi for the academic support provided. This article was developed using the Item Response Theory (IRT) approach with the Graded Response Model (GRM) and is expected to make a positive contribution to the advancement of research in the field of mathematics education.

REFERENCES

[1] C. E. Rahman, Arief Aulia Nasryah, *Evaluasi Pembelajaran*. Ponorogo: Uwais Inspirasi Indonesia, 2019.

[2] M. S. Telaumbanua *et al.*, "Evaluasi dan Penilaian pada Pembelajaran Matematika," *Journal on Education*, vol. 06, no. 01, pp. 4781–4792, 2023.

[3] G. J. Steen, "Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension," *Front Psychol*, vol. 14, Sep. 2023, doi: 10.3389/fpsyg.2023.1242888.

[4] T. Siler, *Think Like a Genius: Use Your Creativity in Ways that Will Enrich Your Life*. Bantam, 1997.

[5] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago: The University of Chicago Press, 1980.

[6] G. J. Steen, "Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension," *Front Psychol*, vol. 14, Sep. 2023, doi: 10.3389/fpsyg.2023.1242888.

[7] D. Özdemir and A. Kınık Topalsan, "Metaphorical Perceptions of Gifted Students towards Mathematics and Science Concepts," *Educational Process International Journal*, vol. 11, no. 3, 2022, doi: 10.22521/edupij.2022.113.6.

[8] O. Thibodi, "Metaphors for Learning Mathematics: An Interpretation Based on Learners' Responses to an Exploratory Questionnaire on Mathematics and Learning," *International Journal of Secondary Education*, vol. 5, no. 6, p. 70, 2017, doi: 10.11648/j.ijsedu.20170506.11.

[9] I. Nurhikmayati, "Pembelajaran dengan Pendekatan Metaphorical Thinking untuk Meningkatkan Kemampuan Pemahaman dan Penalaran Matematis Siswa SMP," UPI, 2013.

[10] K. G. D. Yanti, I. G. N. Pujawan, and G. A. Mahayukti, "Meningkatkan Kemampuan Penalaran Matematis Siswa Melalui Penerapan Pendekatan Metaphorical Thinking," *Jurnal IKA*, vol. 16, no. 2, p. 84, Sep. 2019, doi: 10.23887/ika.v16i2.19828.
[11] D. Andriani and G. Hamdu, "Analisis Rubrik Penilaian Berbasis Education for Sustainable Development dan Konteks Berpikir Sistem di Sekolah Dasar," *EDUKATIF : JURNAL ILMU PENDIDIKAN*, vol. 3, no. 4, pp. 1326–1336, Jun. 2021, doi: 10.31004/edukatif.v3i4.514.
[12] J. H. Nieminen and J. Lahdenperä, "Assessment and epistemic (in)justice: how assessment produces knowledge and knowers," *Teaching in Higher Education*, vol. 29, no. 1, pp. 300–317, Jan. 2024, doi: 10.1080/13562517.2021.1973413.
[13] B. Winter and J. Yoshimi, "Metaphor and the Philosophical Implications of Embodied Mathematics," *Front Psychol*, vol. 11, Nov. 2020, doi: 10.3389/fpsyg.2020.569487.
[14] L. Arriza, H. Retnawati, and R. T. Ayuni, "Item Analysis of High School Specialization Mathematics Exam Questions with Item Response Theory Approach," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 1, pp. 0151–0162, Mar. 2024, doi: 10.30598/barekengvol18iss1pp0151-0162.
[15] B. Baharudin, C. Chairuddin, and T. Tahir, "Pengembangan Lembar Kerja Peserta Didik Berbasis Pendekatan Metaphorical Thinking Terhadap Kemampuan Pemecahan Masalah Matematis," *JSN : Jurnal Sains Natural*, vol. 2, no. 2, pp. 29–34, May 2024, doi: 10.35746/jsn.v2i2.432.
[16] R. Febriano, E. Tandililing, and E. Enawaty, "Pengembangan Instrumen Tes Kemampuan Berpikir Kritis Matematis Dengan Menggunakan Analisis Model Rasch Pada Siswa SMP," *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK)*, vol. 10, no. 9, 2021.
[17] I. Falani, Y. Ramalisa, S. Sainuddin, and Kriswantoro, "Development of a PISA-based mathematical literacy instrument for Indonesian students using item response theory," *J Phys Conf Ser*, vol. 3148, no. 1, p. 012003, Nov. 2025, doi: 10.1088/1742-6596/3148/1/012003.
[18] H. DİLEK and U. AKBAŞ, "Investigation of education value perception scale's psychometric properties according to CTT and IRT," *International Journal of Assessment Tools in Education*, vol. 9, no. 3, pp. 548–564, Sep. 2022, doi: 10.21449/ijate.986530.
[19] R. K. Hambleton, Swaminathan, and J. Rogers, *Fundamentals of Item Response Theory*. london: SAGE Publication inc, 1991.
[20] M. S. Sarea and R. Ruslan, "Karakteristik Butir Soal: Classical Test Theory vs Item Response Theory?," *DIDAKTIKA : Jurnal Kependidikan*, vol. 13, no. 1, pp. 1–16, Aug. 2019, doi: 10.30863/didaktika.v13i1.296.
[21] B. Baharudin, C. Chairuddin, and T. Tahir, "Pengembangan Lembar Kerja Peserta Didik Berbasis Pendekatan Metaphorical Thinking Terhadap Kemampuan Pemecahan Masalah Matematis," *JSN : Jurnal Sains Natural*, vol. 2, no. 2, pp. 29–34, May 2024, doi: 10.35746/jsn.v2i2.432.
[22] F. Fitriani, "Penerapan Pembelajaran Metaphorical Thinking Pada Siswa SMP," *MEGA: Jurnal Pendidikan Matematika*, vol. 1, no. 1, pp. 8–15, Mar. 2020, doi: 10.59098/mega.v1i1.177.
[23] I. Falani, "Desain dan Validasi Aplikasi Tes Literasi Matematika Berbasis Komputer dengan Pendekatan Item Response Theory," *Edu-Sains: Jurnal Pendidikan Matematika dan Ilmu Pengetahuan Alam*, vol. 12, no. 2, pp. 16–28, Jul. 2023, doi: 10.22437/jmpmipa.v12i2.26716.
[24] N. Aini, A. Sari, V. Antia, U. S. Daimah, I. Muhakimah, and S. S. Dewanti, "Konstruksi Instrumen Tes Kemampuan Pemecahan Masalah Menggunakan Teori Respon Butir," vol. 09, no. September, pp. 193–206, 2024.
[25] S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara, 2013.
[26] L. L. Oreondo, *Evaluating Educational Outcomes*. Rex Bookstore, 2005.
[27] L. R. Aiken, "Three Coefficients for Analyzing the Reliability and Validity of Ratings," *Educ Psychol Meas*, vol. 45, no. 1, pp. 131–142, Mar. 1985, doi: 10.1177/0013164485451012.