PROSIDING DISKUSI PANEL NASIONAL PENDIDIKAN MATEMATIKA

Original Research

Analysis of Assessment Reasoning Mathematics Ability and Curiosity Character Secondary Students' Through Winstep Application

Rahmah Meylina Putri [1], R. Nayla Amanda[2], Riki Amperadi[3], Ayu Faradillah[4)*], & Asih Miatun[5]

[1,2,3,4,5]Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia

INFO ARTICLES

Key Words:

Mathematical reasoning skills; Curiosity character; Quantitative research; Survey approach; Rasch analysis

Abstract: *The problem of students' low mathematical reasoning skills is still a common problem in learning mathematics, many students find it difficult to analyze information and solve non-routine problems. The problem is further exacerbated by a limited curiosity, as evidenced by students' lack of interest in exploring mathematical concepts on their own. Therefore, there is a strong need for valid and reliable assessment tools to measure such cognitive and affective aspects. So, this study investigated the mathematical reasoning ability of junior high school students and mapped their curiosity character. The design of this study was a quantitative survey involving 269 students in the eighth and ninth grades from three schools in Jakarta and Bekasi. Data collection was carried out using a mathematical reasoning ability test of 8 items based on NCTM standards and a curiosity questionnaire of 16 items and analyzed using the Rasch Model. The results confirmed strong psychometric properties for both instruments, with item reliability of 0.99 and unidimensionality greater than 85%, indicating that they are stable, fair, and generalizable to educational assessment.*

Abstrak: Rendahnya kemampuan penalaran matematis siswa dan terbatasnya rasa ingin tahu menjadi tantangan dalam pembelajaran matematika, terutama untuk menyelesaikan masalah non-rutin. Oleh karena itu, instrumen penilaian yang valid dan reliabel sangat dibutuhkan untuk mengukur kedua aspek kognitif dan afektif tersebut secara akurat. Penelitian kuantitatif dengan desain survei ini menguji kemampuan penalaran matematis sekaligus memetakan karakter rasa ingin tahu siswa SMP. Melibatkan 269 siswa kelas VIII dan IX dari tiga sekolah di Jakarta dan Bekasi, data dikumpulkan menggunakan tes penalaran matematis berbasis standar NCTM (8 butir) dan kuesioner rasa tngin tahu (16 butir). Data selanjutnya dianalisis menggunakan Model Rasch. Hasil analisis mengonfirmasi kualitas psikometrik yang kuat dari kedua instrumen, dengan reliabilitas butir mencapai 0,99 dan unidimensionalitas melebihi 85%. Temuan ini menjamin stabilitas, keadilan, serta kelayakan instrumen untuk digunakan dalam penilaian pendidikan matematika.

Correspondence Address: Jl. Tanah Merdeka No.20, RT.11/RW.2, Rambutan, Kec. Ciracas, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta 13830, Indonesia; e-mail: ayufaradillah@uhamka.ac.id

How to Cite (APA 6th Style): Putri, R. M., Amanda, R. N., Amperadi, R., Faradillah, A., & Miatun, A. (2026). Analysis of Assessment Reasoning Mathematics Ability and Curiosity Character Secondary Students' Through Winstep Application. *Prosiding Diskusi Panel Nasional Pendidikan Matematika*, 69-84.

INTRODUCTION

Mathematical reasoning ability is a basic cognitive competence that students in secondary education need to develop in order to solve increasingly complex mathematical problems in the twenty-first century. Given the importance of critical thinking, problem solving, and data-driven decision making in education, reasoning has become a vital part of learning mathematics (Hamidah et al., 2025; Zhang et al., 2025). Reasoning and proof have been identified as core process standards in mathematics learning, highlighting the fact that this competency is a connecting thread across mathematical concepts (Valenta et al., 2024). In particular, Tum (2024) describes mathematical reasoning as a logical process of mathematical communication and thinking by which learners analyze conceptual relationships, build premises, and reach valid conclusions. Hu et al. (2025) also argue that reasoning involves rigorous application of premises to draw valid conclusions, but also requires students to critically evaluate arguments in the face of unfamiliar circumstances. This cognitive capacity has theoretical importance, but empirical evidence shows that students' mathematical reasoning performance is still unsatisfactory in many learning environments (Oktaviani & Hanifah, 2025). Many students still have major difficulties when faced with non-routine problems that require analytical and logical thinking beyond the basic procedural knowledge (Tamaulina, 2025). This continuous deficiency of mathematical reasoning is an urgent educational problem that calls for timely attention and accurate assessment (Li et al., 2023) as the mastery of these skills is necessary for higher academic performance.

While mathematical reasoning represents a crucial cognitive competency, its development is closely influenced by students' affective characteristics, particularly curiosity. Curiosity has been recognized as a powerful motivational force that stimulates cognitive engagement and knowledge construction. According to Poli et al. (2024), curiosity reflects an internal drive that motivates learners to seek, investigate, and acquire new information to satisfy gaps in their understanding. In a similar vein, Forss et al. (2024) emphasize that curiosity enables learners to explore challenges and tolerate uncertainty in complex learning situations. In mathematics learning, curiosity can therefore be conceptualized as an affective disposition that encourages students to actively investigate concepts, explore alternative solution strategies, and construct a deeper understanding.

However, the above limitations of mathematical reasoning are often met with a very low level of curiosity in learners (Oktaviani & Hanifah, 2025). The challenges of solving complex mathematical problems are highly related to the limited exploratory behaviors and the reluctance to independently explore unfamiliar mathematical situations (Li, Zhang, & An, 2023). These two aspects are interrelated. Students who are not curious tend to be less engaged in mathematical exploration, which in turn impedes the development and application of mathematical reasoning in problem solving. It is therefore equally important to tackle affective issues, such as low levels of curiosity, as it is to address cognitive shortcomings in reasoning.

Valid and reliable measures are absolutely critical to accurately assess and manage these complex cognitive and affective problems. Thus, Rasch Model is a good psychometric solution for assessing such instruments (Dwi & Faradillah, 2025). Winsteps has been widely applied in research on Indonesian mathematics learning, because it enables the researchers to estimate item difficulty and person ability on a single logit scale, analyze rating scale functioning through Wright maps, and identify potential bias through Differential Item Functioning (DIF) analysis (Faradillah & Adlina, 2021; Faradillah & Febriani, 2021). For instance, Syahwa et al. (2024) demonstrated that Rasch analysis based on Winsteps always produces a stable item hierarchy and high reliability indices. Hence, this method is a highly appropriate and powerful tool to robustly assess the cognitive and affective constructs.

Despite the availability of advanced psychometric tools like Winsteps, previous educational research has largely overlooked the measurement quality of mathematical reasoning instruments and curiosity. Although numerous previous studies have explored mathematical reasoning and curiosity extensively (Alsalamah, 2022; Bernard et al., 2025; Fauziah & Hendriana, 2022; Fatin et al., 2024; Muharni et al., 2024; Rahmah et al., 2024), they have predominantly focused on instructional interventions and correlational relationships rather than the structural validity of the instruments themselves. In many cases, instrument validation remains overly dependent on Classical Test Theory (CTT) approaches, which provide limited information regarding item functioning, respondent ability, and potential measurement bias (Fischer et al., 2021). Consequently, empirical evidence concerning the simultaneous psychometric evaluation of paired cognitive and affective assessment instruments remains insufficient, revealing a clear gap in the mathematics learning literature.

The present research closes the above gap by changing the analytical perspective from the classical statistics to a modern item response approach. The application of this methodology to evaluate both instruments overcomes the limitations of student-sample dependence and provides an objective calibration of item quality. The novelty of this research is that it performs a simultaneous Rasch-based psychometric evaluation of the mathematical reasoning test and the curiosity questionnaire in a single analytical framework. In this study in particular, the aim is to examine the psychometric properties of the mathematical reasoning test and the curiosity questionnaire by applying the Rasch Model in terms of item and person reliability, the functioning of a rating scale, unidimensionality, and Differential Item Functioning (DIF).

METHOD

Research design, site, and participants

This study employs a quantitative method using a survey approach. The quantitative approach was chosen because the study focuses on objectively measuring variables through numerical data analyzed using statistical methods. Meanwhile, the survey approach was used to gather information regarding the characteristics, beliefs, behaviors, and responses of students to the variables under investigation. Survey research serves as an efficient approach to collect structured and standardized data from a relatively large population regarding respondents' currently occurring or past opinions, beliefs, characteristics, or actions (Syahwa et al., 2024).

This study was conducted at three junior high schools located in two different regions, namely Bekasi and Jakarta. The research subjects were selected using purposive sampling. Purposive sampling involves selecting participants based on deliberate criteria that align directly with the research goals and the specific traits needed for the study (Tajik et al., 2024). This technique was chosen because the selection of classes was conducted randomly by the mathematics teachers based on the availability of classes that could be used as research subjects. Three classes were selected from each school, resulting in a total of nine classes involved in the study.

The characteristics of respondents in this study are age, grade level and school of origin. The demographic profile of the respondents is shown in Figure 1. The demographic data in Figure 1 shows that the sample is perfectly balanced at 1:1 for age and gender. But the study was largely dominated by Grade IX students and Jakarta residents in an approximate 2:1 ratio to their counterparts. Specifically, participants who were grade IX students comprised 64.31% of the participants, and residents from Jakarta comprised almost the same majority at 64.68%.

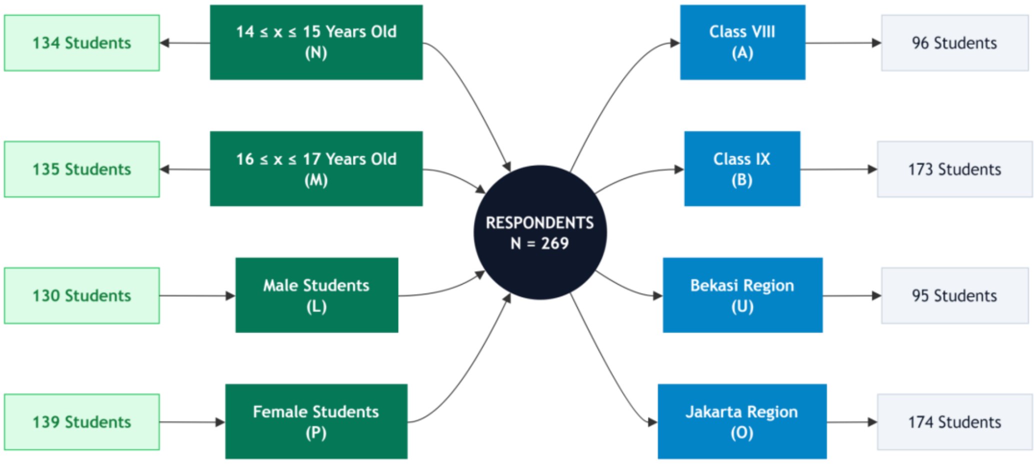

Figure 1. Tabulation Data

Instrument

The current study focuses on evaluating the psychometric properties of the instruments used to measure secondary students' mathematical reasoning ability and curiosity character. This study adopted and expanded a test instrument aimed to examine the level of students' mathematical reasoning abilities on the Pythagorean theorem topic in Indonesia (Novianda et al., 2021). Novianda offered a mathematical reasoning assessment framework grounded by the National Council of Teachers of Mathematics (NCTM) reasoning and proof standards. The eight test items were developed using open-ended essay questions divided into four distinct scales: (1) mathematical manipulation, (2) draw conclusions and compile evidence, (3) formulate and assess argument validation, and (4) perform generalization (Novianda et al., 2021). Evaluation of the students' answers was calculated based on a five-point scoring rubric ranging from score 0 for incorrect or missing responses to score 4 for a correct and complete answer.

For the second aspect, a questionnaire measuring students' curiosity character within the mathematics learning environment was adopted and expanded (Nambiar et al., 2022). Nambiar et al. offered a curiosity assessment framework grounded by psychological and behavioral indicators of curiosity contextualized within mathematics learning. The sixteen questionnaire items are distributed evenly with four items allocated to assess each of the four specified dimensions: (1) pay attention, (2) take notes, (3) asking, and (4) comparing (Nambiar et al., 2022). Assessment criteria for these behavioral indicators utilize a four-point measurement scale based on a frequency response format. Scoring for all items ranges from a minimum score of never (1) to always (4).

Translation into Bahasa

The original instrument was translated into Bahasa Indonesia to ensure that the secondary school participants could comprehend the items effortlessly and find them relevant to their experiences. This translated version was subsequently reviewed and verified by bilingual scholars to maintain cross-cultural adaptation and instrument validity (Mulyono et al., 2023).

Data analysis

The data collected from both the paper-based tests and questionnaires were compiled, scored, and tabulated using a predetermined coding system. The Rasch Model was employed as the primary analytical framework to evaluate the psychometric properties of both instruments. This analysis was executed through the Winsteps application, which required converting the dataset into formatted text prior to processing (Mulyono et al., 2023).

The evaluation process involved two distinct phases. The first phase focused on data screening to identify missing responses and eliminate person outliers based on the outfit Mean Square (MNSQ) criteria (Nizaruddin et al., 2024; Febryana et al., 2024). Any data identified as an outlier was removed from the dataset to ensure measurement cleanliness.

The second phase examined the overall quality of the remaining cleaned data through a comprehensive psychometric evaluation (Dhawo et al., 2026). This phase included several specific analyses: (a) dimensionality to ensure unidimensionality; (b) item and person separation and reliability; (c) rating scale functioning; and (d) item fit validation based on the acceptable outfit Mean Square (MNSQ), Z-Standard (ZSTD), and Point Measure Correlation (PT-Corr) values (Mulyono et al., 2023; Syahwa et al., 2024; Febryana et al., 2024).

RESULT

Unidimensionality analysis of items

The psychometric evaluation was initiated by examining the measurement characteristics of mathematical reasoning ability test and curiosity questionnaire using the Rasch Model. This analysis aimed to determine whether both instruments satisfied the fundamental measurement requirements before interpreting students' cognitive and affective profiles. According to Rasch measurement principles, item reliability, person reliability, separation indices, and model fit provide important evidence regarding the quality and consistency of educational assessment instruments (Faradillah & Adlina, 2021; Syahwa et al., 2024). The summary of these psychometric indicators is presented in Table 1.

Table 1. Summary of Rasch measurement model on Curiosity and Mathematical Reasoning Ability constructs

Mathematical Reasoning Ability						
Parameter (with quality criteria)		**Global Scale (8 item)**	**G* (2 items)**	**VA* (2 items)**	**DC* (2 items)**	**MP* (2 items)**
Model fit: Summary of items						
Item	mean in logit	.00, SD = 1.09	.00, SD = 2.29	.00, SD = .12	.00, SD = .89	.00, SD = .54
	reliability	.99	.99	.00	.99	.97
	separation reliability	.99	.99	.00	.99	.97
	model fit MNSQ range extremes	Infit				
		.62 - 1.16	.71-.81	1.00-1.01	.79-.90	.88-1.06
		Outfit				
		.41 - 3.52	.58-1.66	.91-.93	1.00-1.02	.81-1.02
	separation index	13.01	10.20	.00	8.63	5.30
Separate item strata		17.68 = 18 level	13.93 = 14 level	.33 = 0 level	11.84 = 12 level	7.40 = 7 level
Model fit: Summary of persons			3.68, SD = 3.34	2.45, SD = 3.01	-1.38, SD = 2.27	1.04, SD = 2.23

Person	means in logits	-1.01, SD = 1.54	-3.68, SD = 3.34	-2.45, SD = 3.01	-1.38, SD = 2.27	1.04, SD = 2.23
	reliability	.84	.71	.70	.59	.60
	separation reliability	.84	.71	.70	.59	.60
	separation index	2.30	1.58	1.53	1.21	1.21
Separate Person strata		3.40 = 3 level	2.44 = 2 level	2.37 = 2 level	1.94 = 2 level	1.94 = 2 level
				Curiosity Characters		
Parameter **(with quality criteria)**		**Global Scale (8 item)**	**Co* (4 items)**	**As* (4 items)**	**TN * (4 items)**	**PA* (4 items)**
Model fit: Summary of items						
Item	mean in logit	.00, SD = 0.83	.00, SD = 0.60	.00, SD = 0.73	.00, SD = 1.24	.00, SD = 0.64
	reliability	.99	.97	.97	.99	.95
	separation reliability	.99	.97	.97	.99	.95
	model fit MNSQ range extremes			Infit:		
		.76 - 1.56	.88-1.13	.85-1.36	.88-1.21	.80-1.14
				Outfit		
		.64 - 1.60	.89-1.14	.84-1.41	.87-1.20	.79-1.13
	separation index	8.21	5.56	6.23	11.10	4.33
Separate item strata		11.28 = 11 level	7.74 = 8 level	8.64 = 9 level	15.13 = 15 level	6.10 = 6 level
Model fit: Summary of persons						
Person	means in logits	-1.01, SD = 1.54	1.29, SD = 1.78	1.66, SD = 1.81	1.82, SD = 1.67	3.23, SD = 1.86
	reliability	.81	.64	.64	.55	.42
	separation reliability	.84	.64	.64	.55	.42
	separation index	2.07	1.35	1.33	1.11	.85
Separate Person strata		3.09 = 3 level	2.13 = 2 level	2.10 = 2 level	1.81 = 2 level	1.46 = 1 level

*Note. The abbreviations denote the specific indicators of the two measured variables. Mathematical Reasoning constructs: G = Generalization; VA = Validating Arguments; DC = Drawing Conclusions and compiling evidence; MP = Mathematical Manipulation. Curiosity characters: Co = Comparing; As = Asking; TN = Taking Notes; PA = Paying Attention.

Based on Table 1, both instruments generally demonstrated satisfactory psychometric quality. Item reliability values ranged from .95 to .99 across most constructs, while item separation indices exceeded the recommended criterion of 3.00 for nearly all dimensions. These findings indicate that the calibrated items were distributed across several distinguishable levels of difficulty, allowing the instruments to differentiate between relatively easy and more difficult items (Bintang & Suprananto, 2024). In addition, the average item measures were centred at approximately .00 logits, indicating that the calibration process followed the standard Rasch measurement procedure.

Compared with the item statistics, person reliability values were relatively lower, ranging from .42 to .84 across the reasoning and curiosity instruments. Similarly, person separation indices generally classified respondents into one or two distinguishable ability strata. Nevertheless, the Infit and MNSQ statistics for both persons and items remained predominantly within the recommended Rasch criteria, indicating that the observed response patterns were generally consistent with model expectations (Faradillah & Adlina, 2021; Syahwa et al., 2024).

An exception was found for the VA indicator which produced an item reliability and separation index of .00. This outcome was characterized by a very small standard deviation of item difficulty (SD = 0.12 logits), showing little variability among the calibrated items within this indicator. The low variation in item difficulty can reduce the value of the separation index, which is based on Rasch measurement theory, because the items have almost the same position on the latent continuum (Bintang & Suprananto, 2024). In general, the psychometric evidence provided in Table 1 indicated that both instruments satisfied the basic measurement criteria for the subsequent analyses, including rating scale functioning, Wright maps interpretation, and DIF.

Effectiveness of the rating scale

To further evaluate the quality of the two instruments, the functionality of the rating scales was investigated with the Rasch Rating Scale Model . This analysis was designed to test that each response category operated sequentially and represented increasing levels of the latent traits being measured. Rasch measurement principles state that a good rating scale must have ordered average measures, progressive Andrich threshold, and acceptable MNSQ values (Faradillah & Adlina, 2021). The functioning statistics for the mathematical reasoning test and the curiosity questionnaire are presented in Table 2 for the category.

Table 2. Summary of category structure of rating scales model on Curiosity and Mathematical Reasoning Ability

Rating scale step number	Observed person	Average calibration	Outfit MNSQ	Threshold	Threshold distance
Mathematical Reasoning					
0	429 (80%)	-4.23	.60	NONE	(-3.28)
1	13 (2%)	-2.07	.30	-1.91	1.37
2	18 (3%)	.46	1.58	-1.24	0.67
3	26 (5%)	1.64	1.82	.29	1.53
4	52 (10%)	4.06	1.19	2.86	2.57
Curiosity					
Rating scale step number	**Observed person**	**Average calibration**	**Outfit MNSQ**	**Threshold**	**Threshold distance**
1	32 (3%)	-1.62	1.04	NONE	(-3.99)
2	272 (25%)	-.01	1.02	-2.86	1.56
3	409 (38%)	1.46	1.01	.30	3.16
4	363 (34%)	2.84	1.00	2.56	2.26

Table 2 indicates that the rating scales of both instruments were performing well, with the average calibration values and Andrich thresholds increasing monotonically across the categories. There were no threshold reversals. Calibrations ranged from –4.23 to 4.06 logits for the mathematical reasoning test and from –1.62 to 2.84 logits for the curiosity questionnaire. Although the reasoning test exhibited minor Outfit MNSQ deviations, both scales remained ordered and interpretable

(Faradillah & Adlina, 2021). These results confirmed that the scoring categories were consistent and appropriate for further Rasch analysis (Faradillah & Adlina, 2021; Syahwa et al., 2024).

Item and Person mapping

Following the psychometric evaluation and rating scale analysis, the relationship between students' ability and item difficulty was further examined using the Wright maps. The Wright maps simultaneously displays person measures and item calibrations on a common logit scale, enabling direct comparison between respondents' mathematical reasoning ability or curiosity level and the relative difficulty of each assessment item (Faradillah & Adlina, 2021). The person item distributions for the curiosity questionnaire and the mathematical reasoning test are presented in Figures 2 and 3, respectively.

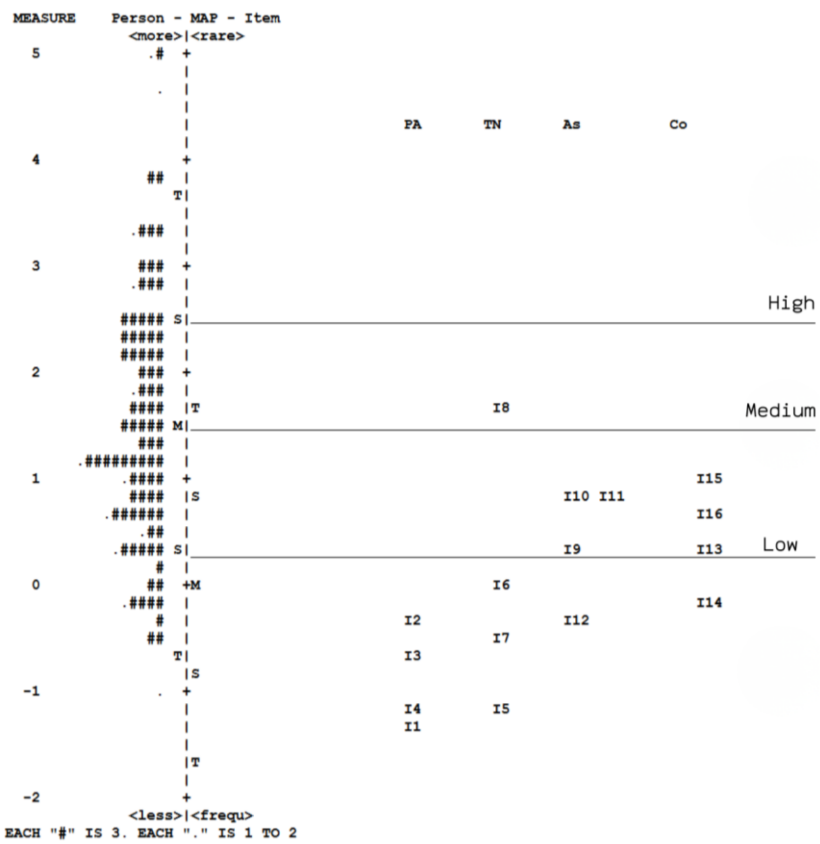

Figure 2. Item and Person Mapping for Curiosity

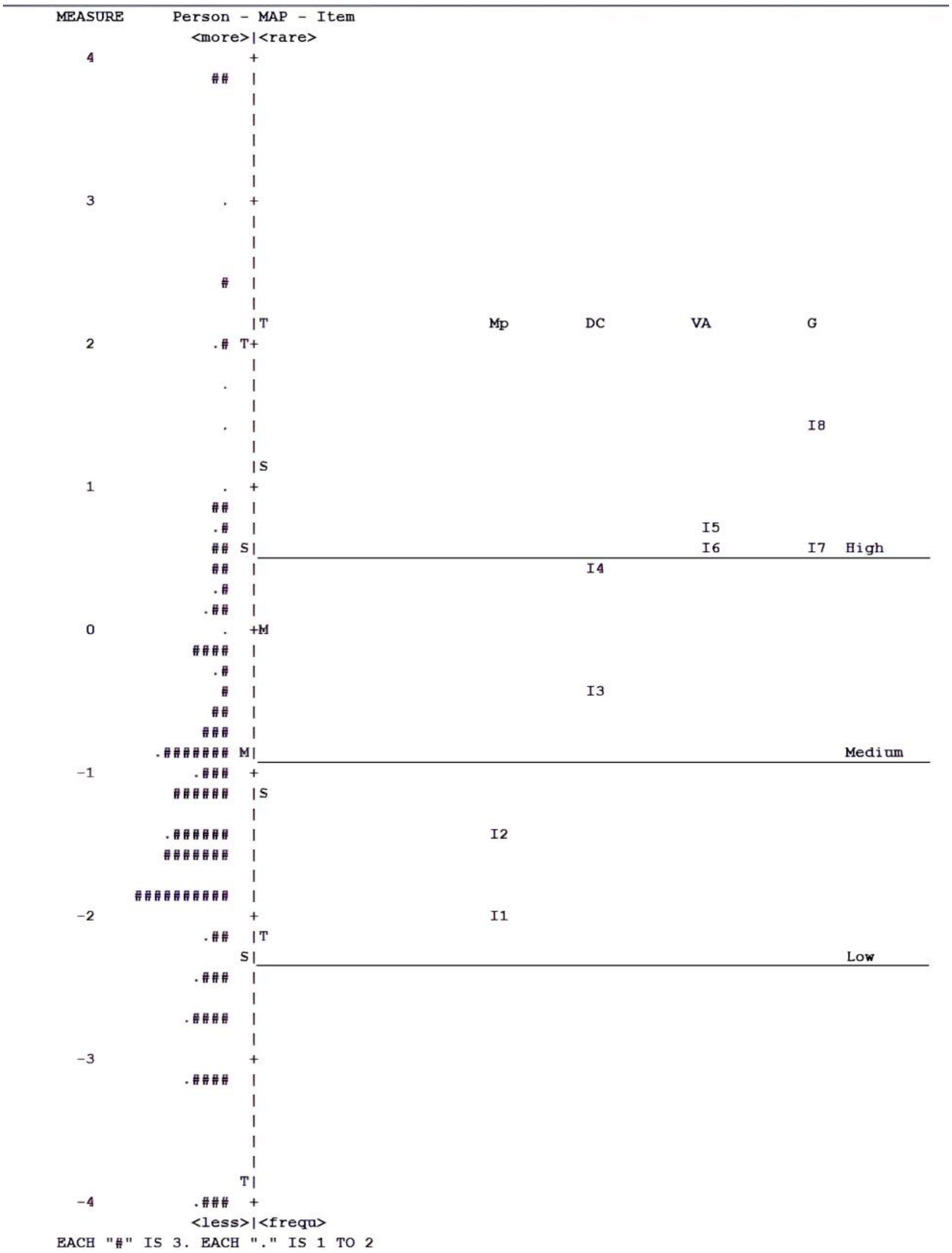

Figure 3. Item and Person Mapping for Mathematical Reasoning

Figures 2 and 3 present the Wright maps illustrating the distribution of person measures and item difficulty for the curiosity questionnaire and the mathematical reasoning test, respectively. In both instruments, most respondents were concentrated around the middle to upper region of the logit continuum, indicating that the majority of students demonstrated moderate to relatively high levels of curiosity and mathematical reasoning. At the same time, the distribution of item measures covered a broad range of difficulty, suggesting that the instruments were able to assess students across different levels of the measured constructs (Faradillah & Adlina, 2021; Syahwa et al., 2024).

The most difficult item in the curiosity questionnaire is Item I15 ("I compare the mathematics teacher's explanations with other sources regarding a topic"), followed by Items I10, I11 and I16, which all represent behaviours involving questioning and comparing mathematical ideas . High logit positions indicate that students were less likely to engage in independent inquiry, critical comparison and active classroom participation than other curiosity-related behaviours. Conversely, Item I1 ("I pay attention when the teacher explains mathematical materials") and Item I4 ("I pay attention to each mathematical assignment given by the teacher") were at the lower end of the logit scale, suggesting that most students could easily exhibit teacher-directed learning behaviors.

A similar pattern emerged in the mathematical reasoning test. Item I8, which required students to construct a general mathematical model connecting the lengths of two diagonals, together with Items I5 and I6 involving argument validation and mathematical justification, occupied the highest logit positions, indicating that higher-order reasoning processes were the most demanding for students. Conversely, Item I1, which required routine application of the Pythagorean Theorem to determine the height reached by a ladder, was calibrated as the easiest item, suggesting that procedural mathematical manipulation was relatively well mastered by the participants. Overall, the Wright maps indicate that students demonstrated stronger procedural reasoning than higher-order reasoning, while their curiosity was more evident in teacher-directed learning behaviours than in inquiry-oriented activities. These findings provide substantive evidence that the calibrated items successfully distinguished different levels of students' mathematical reasoning and curiosity while highlighting specific aspects that require greater instructional attention (Hamidah et al., 2025; Hu et al., 2025; Bintang & Suprananto, 2024).

Item Bias Distributions

To examine the fairness of both assessment instruments across different respondent groups, DIF analysis was performed using the Rasch Model. DIF analysis evaluates whether respondents with comparable levels of the latent trait have different probabilities of responding to an item due to group membership rather than the construct being measured (Faradillah & Adlina, 2021; Syahwa et al., 2024). The DIF statistics for the mathematical reasoning test and the curiosity questionnaire are summarised in Table 3.

Table 3. Item bias model on Curiosity and Mathematical Reasoning Ability

Num	Item	Status*	DIF Measure	DIF Contrast	t	Probability
Item Bias Demography Status (Mathematical Reasoning)						
1	I1	A	-1.21	1.07	6.46	.0000
		B	-2.28			
		N	-1.21	.89	4.67	.0000
		M	-2.10			
2	I5	A	.41	-1.16	-3.69	.0003
		B	1.57			
		N	.41	-.54	-2.53	.0124
		M	.95			
3	I6	A	.37	-.83	-3.01	.0031
		B	1.20			
Item Bias Demography Status (Curiosity)						
1	I3	L	-1.05	-.56	-2.04	.0430
		P	-0.49			
2	I9	A	.59	.52	2.29	.0237
		N	.07			
3	I11	L	.65	-.50	-2.23	.0270
		P	1.15			
4	I13	L	.47	.53	2.26	.0251
		P	-.06			

*Note: The demographic groups of respondents in this study were categorized by age range (N for $14 \leq x \leq 15$ years; M for $16 \leq x \leq 17$ years), gender (L for male; P for female), grade level (A for VIII grade; B for IX grade).

Table 3 presents the DIF analysis conducted to evaluate item invariance across demographic groups, including grade level, age, and gender. Following current recommendations for Rasch-based DIF analysis, items were classified as exhibiting DIF when the absolute DIF contrast exceeded .50 logits with p < .05, indicating potential differences in item functioning after controlling for respondents' ability levels (Humphry & Montuoro, 2021; Pohl et al., 2021). Overall, only a small proportion of items demonstrated significant DIF, suggesting that most items functioned consistently across the demographic groups examined.

For the mathematical reasoning test, three items (I1, I5, and I6) exhibited significant DIF. Specifically, Item I1 favored Grade VIII students over Grade IX students (DIF contrast = 1.07 logits, $p < .001$). Likewise, a similar pattern was observed between the younger and older age groups, with a DIF contrast of .89 logits ($p < .001$). In contrast, Items I5 and I6 favored Grade IX students, yielding DIF contrasts of –1.16 and –0.83 logits, respectively ($p < .01$). These findings indicate that differential functioning in the reasoning test was primarily associated with students' educational level and age.

A comparable pattern was observed in the curiosity questionnaire, where four items (I3, I9, I11, and I13) demonstrated significant DIF. Items I3 and I11 were more readily endorsed by female students, whereas Item I13 favored male students. Furthermore, Item I9 favored Grade VIII students over Grade IX students (DIF contrast = .52 logits, $p = .024$). Taken together, these findings indicate that both instruments generally maintained satisfactory measurement invariance across the examined demographic groups, although the flagged items warrant further investigation in subsequent validation studies to ensure equitable measurement across diverse populations (Humphry & Montuoro, 2021).

DISCUSSION

The current study suggests that both the mathematical reasoning test and the curiosity questionnaire have good psychometric properties for educational assessment. The high item reliability and item separation indices suggest that calibrated items are good indicators of different levels of mathematical reasoning and curiosity, hence providing a stable measurement across respondents. The person reliability was relatively lower but this finding is more likely to be related to the relatively homogeneous nature of the participants than to the weaknesses in the instruments themselves. Likewise, the small amount of variability in the Validating Arguments indicator suggests that adding more items with different levels of difficulty would improve measurement precision. In summary, the findings of this study confirm that both instruments are suitable for measuring students' cognitive and affective characteristics (Faradillah & Adlina, 2021; Bintang & Suprananto, 2024; Valenta et al., 2024).

The analysis of the rating scale also confirms the functionality of the response categories under the Rasch model. The ordered pattern of the average measures and the Andrich thresholds suggest that students were generally able to differentiate each response category according to increasing levels of mathematical reasoning and curiosity. While some Outfit MNSQ values exhibited small deviations, they did not disturb the general structure of the categories since the thresholds remained in order. Therefore, the scoring system was able to identify significant differences in the responses of the students and gave dependable estimates for the next Wright maps analysis (Faradillah & Adlina, 2021; Syahwa et al., 2024).

The Wright maps also reveal meaningful differences in students' curiosity profiles. Behaviors related to paying attention and taking notes were easier than behaviors related to questioning and comparing mathematical ideas. The most difficult curiosity behaviour was identified as "I compare the explanations of the mathematics teacher with other sources on a topic" indicating that students were less used to independently evaluating mathematical information. This finding indicates that classroom learning is still mainly teacher-directed, with little scope for students to develop inquiry, comparison and critical reflection. Similar results were found by Hariyanti and Lestari (2023) and Mayang et al. (2024), who stated that inquiry-oriented learning environments are more conducive for students' curiosity.

A similar pattern was found in the test of mathematical reasoning. Mathematical manipulation was the easiest reasoning process while generalisation and argument validation were much more demanding. This hierarchy suggests that students were generally able to carry out procedural tasks but had more difficulty when asked to justify mathematical arguments or formulate general mathematical relationships. These findings are in line with recent studies that demonstrate higher-order reasoning involves abstraction, logical justification, and conceptual transfer beyond routine procedural skills (Hamidah et al., 2025; Hu et al., 2025). Thus, teaching mathematics should offer more opportunities for students to construct arguments, justify solutions and generalise mathematical concepts through meaningful problem solving activities.

Another result of note relates to person–item targeting and DIF. The Wright maps showed that the instruments were generally aligned with the ability levels of the participants, although several students were above the most difficult items, indicating that more difficult items are needed to improve measurement precision for high-performing students. The DIF analysis revealed a small number of items that operated differently for groups of respondents. These differences may reflect variation in classroom experiences, instructional practices, and opportunities to engage in mathematical discussion rather than invalid measurement. Hence, further research should improve the item difficulty and study the possible influence of the context on the performance of the items (Bintang & Suprananto, 2024; Syahwa et al., 2024).

In general, the results indicate that the Rasch Model offers evidence for the quality of the instrument and meaningful insights into students' mathematical reasoning and curiosity profiles. From an educational point of view, the results point to the importance of strengthening learning environments that promote questioning, comparison, argumentation and mathematical generalisation, rather than focusing solely on procedural competence. At the same time, from a psychometric perspective, increasing the item difficulty range and improving the person–item targeting will further enhance the accuracy, fairness, and diagnostic value of both instruments in future mathematics assessment.

CONCLUSION

The study successfully mapped the profiles of 269 secondary students using the Rasch Model and assessed the psychometric properties of the mathematical reasoning test and the curiosity characters questionnaire. The results showed that both instruments have good psychometric quality with excellent item reliability (.99), unidimensionality (> 85%), ordered rating scale function and no or minimal bias of the demographic groups based on DIF analysis. Furthermore, in terms of the objective of mapping student profiles, the Wright maps analysis showed that higher-order reasoning processes such as generalization and argument validation (I5, I6, I7, and I8), and independent curiosity behaviors such as questioning and comparing (I10, I11, I15, and I16) are significantly more challenging for students than procedural processes (I1) and teacher-directed behaviors such as paying attention and taking notes (I1, I4, and I5). DIF analysis suggests that both instruments have acceptable measurement invariance across grade, age and gender, though a few flagged items (three in reasoning and four in curiosity) show minor group differences. These empirical findings are valid assessment tools but also point to the need for strategies to promote complex reasoning, although the flagged DIF items need some minor refinement in future studies. These empirical findings ultimately provide valid and reliable assessment tools for mathematics education while highlighting the critical need for instructional strategies that actively promote complex reasoning and independent inquiry.

ACKNOWLEDGMENT

The authors would like to thank the principals, mathematics teachers, and students of secondary schools involved in Jakarta and Bekasi for administrative permission and facilitation of data collection. We are also very grateful to the expert validators for their rigorous feedback and constructive guidance that greatly improved the validity of the instruments. Finally, the authors would like to express special thanks to the lecturers and colleagues of the Evaluation of Learning Processes and Outcomes course at Universitas Muhammadiyah Prof. DR. HAMKA. The Academic support, joint discussions and continued guidance of theirs was very vital for the successful completion and final publication of this research paper.

REFERENCE

Alsalamah, N. A. (2022). Kemampuan penalaran matematis dalam pembelajaran flipped classroom bernuansa etnomatematika berbantuan Edpuzzle. ANARGYA: Jurnal Ilmiah Pendidikan Matematika, 5(1). https://doi.org/10.24176/anargya.v5i1.7578

Bernard, B., Sahid, S., & Mutmainnah, F. (2025). Analisis kemampuan penalaran matematis dalam memecahkan masalah dimensi tiga ditinjau dari self-regulated learning. Pedagogy: Jurnal Pendidikan Matematika, 10(2), 653–666.https://doi.org/10.30605/pedagogy.v10i2.6174

Bintang, R. S., & Suprananto, S. (2024). The impact of sample size, test length, and person-item targeting on the separation reliability in Rasch model: A simulation study. JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 13(1), 1–13. https://doi.org/10.15408/jp3i.v13i1.27975

Dhawo, M. S., Dias, M. F. A. A., & Rifa'at, M. (2026). Psychometric evaluation of the indonesian version of the holistic nursing competence scale (HNCS). *Journal Ners*, *21*(2), 224–234. https://doi.org/10.20473/jn.v21i2.81471

Dwi, R., & Faradillah, A. (2025). Evaluation of the Difficulty of Self-Efficacy Items in Mathematics Learning Using Rasch Model Analysis. *SJME*, *09*(02), 357–369. https://doi.org/10.35706/sjme.v9i2.200

Faradillah, A., & Adlina, S. (2021). Validity of critical thinking skills instrument on prospective mathematics teachers. Jurnal Penelitian dan Evaluasi Pendidikan, 25(2), 126–137. https://doi.org/10.21831/pep.v25i2.40662

Faradillah, A., & Febriani, L. (2021). Mathematical trauma students' junior high school based on grade and gender. *Infinity*, *10*(1), 53–68. https://doi.org/10.22460/infinity.v10i1.p53-68

Fatin, N., Zaenuri, Z., & Walid, W. (2024). Kemampuan bepikir kritis matematis ditinjau dari rasa ingin tahu dalam model pembelajaran PBL dengan pendekatan kontekstual. EDUKATIF: Jurnal Ilmu Pendidikan, 6(1), 198–209. https://doi.org/10.31004/edukatif.v6i1.5874

Fauziah, I. A., & Hendriana, B. (2022). Analisis Kemampuan Penalaran Matematis Peserta Didik SMP Ditinjau dari Self Regulated Learning Pada Masa Pademi COVID-19. HISTOGRAM: Jurnal Pendidikan Matematika, 6(2), 82–94. https://doi.org/10.31100/histogram.v6i2.2376

Febryana, U. I., Rahmania, F., Faradillah, A., & Jusra, H. (2024). Evaluation of Validity and Reliability Using Rasch Model and SPSS on Students ' Mathematical Creative Thinking Ability. *Proceeding Konferensi Nasional Matematika (KNM) XXII*, *5*(1), 1–10. https://doi.org/10.30598/ppcst.2024.knmxxii.1-10

Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit. Frontiers in Psychology, 12(July), 1–10. https://doi.org/10.3389/fpsyg.2021.633896

Forss, S., Ciria, A., Clark, F., Galusca, C., Harrison, D., & Lee, S. (2024). A transdisciplinary view on curiosity beyond linguistic humans: animals, infants, and artificial intelligence. Biological Reviews, 99(3), 979–998. https://doi.org/10.1111/brv.13054

Hamidah, D., Susiswo, Susanto, H., Hakim, Z. A., & Osman, S. (2025). Exploring the mathematical reasoning skills across different levels of prospective mathematics teachers: A mixed-methods investigation. Jurnal Pendidikan Matematika, 19(2), 389–412. https://doi.org/10.22342/mej.v19i2.pp389-412

Hariyanti, F., & Lestari, W. (2023). Upaya meningkatkan rasa ingin tahu dan prestasi siswa melalui pendahuluan matematika merupakan pelajaran wajib yang diberikan pada setiap jenjang pendidikan. Penerapan matematika dalam kehidupan sehari-hari, misalnya untuk menyajikan informasi dalam ber. Jurnal Penelitian Pembelajaran Matematika Sekolah (JP2MS), 7(1), 83–94.

Hu, Z., Liu, J., Liu, Z., Liu, Y., Xie, Z., & Song, Y. (2025). RMath: A logic reasoning-focused datasets toward mathematical multistep reasoning tasks. Proceedings of the AAAI Conference on Artificial Intelligence, 39(22), 24104–24112. https://doi.org/10.1609/aaai.v39i22.34585

Humphry S and Montuoro P (2021) The Rasch Model Cannot Reveal Systematic Differential Item Functioning in Single Tests: Subset DIF Analysis as an Alternative Methodology. Front. Educ. 6:742560. doi: 10.3389/feduc.2021.742560

Li, J., Xue, E., Li, C., & He, Y. (2023). Investigating latent interactions between students' affective cognition and learning performance: Meta-analysis of affective and cognitive factors. Behavioral Sciences, 13(7), 555. https://doi.org/10.3390/bs13070555

Nambiar, D., Alex, J., & Pothiyil, D. I. (2022). Development and Validation of Academic Self-regulated Learning Questionnaire (ASLQ). *IJBS: International Journal of Behavioral Sciences*, *16*(2), 89–95. https://doi.org/10.30491/IJBS.2022.321176.1730

Nizaruddin, N., Muhtarom, M., Murtianto, Y. H., & Sutrisno, S. (2024). Examining the self-regulated learning scale using the Rasch model approach. In *Indonesian Journal of Science and Mathematics Education* (Vol. 7, Issue 3, p. 518). Raden Intan State Islamic University of Lampung. https://doi.org/10.24042/ijsme.v7i3.21831

Novianda, D., Darhim, & Prabawanto, S. (2021). Analysis of students' mathematical reasoning ability in geometry through distance learning. *Journal of Physics: Conference Series*, *1882*(1), 1–9. https://doi.org/10.1088/1742-6596/1882/1/012085

Mayang, A. N. A., Karmila, N., & Anwa, W. S. (2024). Pengaruh rasa ingin tahu terhadap hasil belajar matematika siswa asoka. Pendas: Jurnal Ilmiah Pendidikan Dasar, 4(2), 361–366. https://doi.org/10.23969/jp.v10i2.26657

Muharni, F., Anitra, R., & Husna, N. (2024). Pedadidaktika: Jurnal ilmiah pendidikan guru sekolah dasar kemampuan pemecahan masalah matematis ditinjau dari rasa ingin tahu siswa sekolah dasar. Pedadidaktika: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar, 11(1), 55–68. http://ejournal.upi.edu/index.php/pedadidaktika/index

Mulyono, H., Ningsih, S. R. I. K., & Sholeha, B. (2023). Validating the Academic Writing Creativity and Self-Efficacy Scale : A Rasch Model Analysis. *Indonesian Research Journal in Education*, *7*(1), 121–135. https://doi.org/10.22437/irje .v7i1.24789

Oktaviani, M. F., & Hanifah, H. (2025). Analisis karakter curiosity dalam pelajaran matematika pada siswa kelas IX SMP. JPMI (Jurnal Pembelajaran Matematika Inovatif), 8(3), 285–294. https://doi.org/10.22460/jpmi.v8i3.26728

Pohl, S., Schulze, D., & Stets, E. (2021). Partial measurement invariance: extending and evaluating the cluster approach for identifying anchor items. *Applied Psychological Measurement*, *45*(7–8), 477–493. https://doi.org/10.1177/01466216211042809

Poli, F., O'Reilly, J. X., Mars, R. B., & Hunnius, S. (2024). Curiosity and the dynamics of optimal exploration. Trends in Cognitive Sciences, 28(5), 441–453. https://doi.org/10.1016/j.tics.2024.02.001

Rahmah, L. F., Nurfitriani, D. A., & Kusno, K. (2024). Mengukur rasa ingin tahu siswa SMP terhadap pembelajaran matematika studi analitis. Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika, 7(2), 913–919. https://doi.org/10.30605/proximal.v7i2.4011

Syahwa, K. A., Gustia, S. K., Faradillah, A., & Ulfah, S. (2024). Validity and Reliability of Mathematical Reasoning Ability Instruments in High Schools with Winsteps. *Proceeding Konferensi Nasional Matematika (KNM) XXII*, *5*(1), 11–20

Tajik, O., Golzar, J., & Noor, S. (2024). Purposive Sampling. *International Journal of Education and Language Studies*, *2*(2), 1–9. DOI: 10.22034/ijels.2025.490681.1029

Tamaulina. (2025). Developing Critical and Logical Thinking Skills Through the Use of Non-Routine Problems in Mathematics Education. *The Journal of Academic Science*, *2*(9), 1–8. https://doi.org/https://doi.org/10.59613/hhvase06

Tum, A. (2024). Reasoning skills in mathematics teaching: A meta-synthesis on studies conducted in Turkey. International E-Journal of Educational Studies, 8(16), 45–86. https://doi.org/10.31458/iejes.1389681

Valenta, A., Rø, K., & Klock, S. I. (2024). A framework for reasoning in school mathematics: Analyzing the development of mathematical claims. Educational Studies in Mathematics, 116(1), 91–111. https://doi.org/10.1007/s10649-024-10309-5

Zhang, Y., Zhao, B., Jian, M., & Wu, X. (2025). Cognitive diagnostic analysis of mathematics key competencies based on PISA data. PLoS ONE, 20(2). https://doi.org/10.1371/journal.pone.0315539