Jurnal Elemen, 12(2), 375-391, April 2026
https:/doi.org/10.29408/jel.v12i2.34118

e- ISSN: 2442 - 4226

Undergraduate students' mathematical reasoning in numeracy-based tasks: A Rasch model approach

Dewi Hamidah [1] *, Jerhi Wahyu Fernanda [1], Zun Azizul Hakim [2], Galuh Nuril Latifah [3]

[1] Mathematics Education Study Programs, Universitas Islam Negeri Syekh Wasil Kediri, East Java, Indonesia

[2] Psychology Study Programs, Universitas Islam Negeri Sayyid Ali Rahmatullah Tulungagung, East Java, Indonesia

[3] School of Metallurgy and Environments, Central South University, Changsha, China

* Correspondence: dewi.hamidah@uinkediri.ac.id

Abstract

Mathematical reasoning is essential in higher education because it enables students to formulate conjecture, generalize patterns, and justify. This study examined undergraduate students' mathematical reasoning, operationalized through conjecturing, generalizing, and justifying, in two numeracy tasks. A purposive sample of 185 mathematics education students from the first, third and fifth semesters at a State University in Kediri, Indonesia, participated in the study. Responses were scored using an analytic rubric and analyzed with the Rasch model to estimate item difficulty and person ability, evaluate item fit and reliability, and examine Differential Item Functioning (DIF) across gender, Grade Point Average (GPA), and semester level. The results indicated that conjecturing and justifying were the most challenging aspects for students, while evidence of generalizing was relatively limited. Rasch analysis verified the instrument's validity and reliability, with all items satisfying fit requirements, while DIF analysis indicated no significant demographic bias. However, observed patterns suggested that female students tended to be more systematic and accurate, male students were generally more flexible but less consistent, and students with higher GPAs displayed stronger logical reasoning. These findings highlight the importance of instructional strategies that intentionally foster mathematical reasoning and suggest future research involving multiple institutions and longitudinal designs.

Keywords: mathematical reasoning; numeracy problems; Rasch model; undergraduate students

How to cite: Hamidah, D., Fernanda, J. W., Hakim, Z. A., & Latifah, G. N. (2026). Undergraduate students' mathematical reasoning in numeracy-based tasks: A Rasch model approach. *Jurnal Elemen, 12*(2), 375-391. https://doi.org/10.29408/jel.v12i2.34118

Received: 10 February 2026 | Revised: 12 April 2026
Accepted: 12 May 2026 | Published: 14 May 2026

Introduction

Mathematical reasoning is a fundamental competence in mathematics education because it enables learners to formulate conjectures, identify patterns, construct generalizations, and justify conclusions when solving problems. The transition from conjecture to proof is thus a staged, iterative process in which abductive and inductive elements equip learners to formulate and defend claims, while deductive reasoning ultimately establishes necessity and validity (Joachin-Arizmendi et al., 2024). Its importance has been widely recognised in educational frameworks across countries, as it supports both academic learning and practical decision-making in daily life (Andrews-Larson et al., 2021; Herbert & Williams, 2023). At school level, mathematical reasoning contributes to the cultivation of higher-order thinking skills, enhancing learners' intellectual capacities and promoting deeper understanding of mathematics (Jeannotte & Kieran, 2017). In higher education, such reasoning is particularly important because students are expected to engage with increasingly complex and abstract mathematical ideas in a coherent and defensible manner.

Despite its importance, many undergraduate students still experience difficulty in carrying out mathematical reasoning at a level expected in higher education. These difficulties are not limited to obtaining correct answers, but are especially evident when students are required to formulate conjectures, relate specific cases to more general mathematical statements, and provide logically valid justifications for their conclusions. Undergraduate students, in particular, often face difficulties in constructing valid arguments and proofs, which are essential elements of advanced reasoning (Case & Speer, 2021). This challenge is particularly relevant in mathematics teacher education, because prospective teachers are expected not only to solve problems but also to explain and defend mathematical ideas in a clear and mathematically acceptable way. These persistent difficulties indicate the need to examine undergraduate students' reasoning more closely, especially in terms of how they construct conjectures, generalize patterns, and justify solutions. Prior research indicates that students can apply reasoning in applied contexts such as calculus but struggle in more abstract domains, highlighting the challenges of transitioning from empirical reasoning to formal proof (Case & Speer, 2021; Ellis et al., 2022). This disparity underlines the need for more focused research on reasoning at the undergraduate level.

In addition to cognitive challenges, mathematical reasoning may also vary across student characteristics and learning experiences. For example, gender differences have been observed, with male students often characterized by flexibility in thinking, while female students demonstrate systematic and structured approaches (Rokhima et al., 2019; Rubianti et al., 2022). Similar patterns appear in studies on mathematical communication and problem solving, where females exhibit accuracy and adherence to structured procedures, while males tend to explore alternative strategies (Rusdi et al., 2020). Furthermore, academic achievement indicators such as Grade Point Average (GPA) have been shown to correlate positively with reasoning and problem-solving performance (Hasanah et al., 2019; Kadarisma et al., 2019).

Gender and GPA considerations further necessitate equitable assessment practices. Studies show that while male students may approach problems flexibly, female students

demonstrate higher precision and strategy adherence (Rokhima et al., 2019; Rubianti et al., 2022). Higher GPA students tend to employ more organized reasoning, aligning with prior evidence linking achievement to deeper engagement with mathematical principles (Hasanah et al., 2019). Yet, emerging research highlights that psychological and sociocultural factor, such as math anxiety, neuroticism, and academic stereotypes—also mediate performance differences (Li et al., 2022; Lunardon et al., 2022; Wrigley-Asante et al., 2023). This complexity underscores the importance of assessment tools that both capture reasoning accurately and evaluate their fairness across subgroups. In addition, semester level may reflect differences in students' exposure to advanced mathematical content. Therefore, both cognitive and affective dimensions, such as self-regulation, personality, and motivation, may influence reasoning performance (Minnigh & Coyle, 2023; Vos et al., 2023). These variables do not determine reasoning ability on their own, but they are relevant when examining whether an assessment captures reasoning fairly across subgroups.

Although mathematical reasoning has been widely studied at school level, evidence from higher education remains more limited, particularly for reasoning assessed through contextualised open-ended tasks. In this study, numeracy-based tasks are used as the assessment context, not as a separate research variable. Such tasks require students to interpret situations, identify mathematical relationships, and defend their conclusions, thereby providing a suitable context for observing conjecturing, generalizing, and justifying. Numeracy tasks are non-routine problems that demand integration of mathematical knowledge with real-life contexts, encouraging critical orientation and reasoning (Geiger et al., 2015; Purnomo et al., 2022). What remains insufficiently understood is how well undergraduate students perform on these three aspects and whether the instrument used to assess them functions consistently across student groups.

A second gap concerns measurement. Many studies report reasoning scores descriptively, but fewer examine whether the instrument itself functions well at the item level and whether it is fair for different groups of students. This is where Rasch measurement becomes relevant. Compared with Classical Test Theory, which relies strongly on sample-dependent total scores, the Rasch model estimates item difficulty and person ability on the same logit scale, evaluates how well each scored component fits the expected model, and supports Differential Item Functioning analysis to test whether items operate differently across subgroups (Bond et al., 2020; Boone, 2016; Edwards & Alcock, 2010). These features make Rasch analysis especially useful for open-ended reasoning assessments, where researchers need evidence not only of overall scores but also of construct functioning, item hierarchy, and fairness.

A further challenge lies in assessment. Although mathematical reasoning has been widely recognised as an essential learning outcome, many studies still report students' performance in broad terms without showing whether specific reasoning processes are adequately captured. This is problematic because reasoning is multidimensional and cannot be fully represented by final answers alone. In the PISA 2022 mathematics framework, reasoning is positioned within mathematical literacy, defined as the capacity to reason mathematically and to formulate, employ, and interpret mathematics in real-world contexts (OECD, 2022). More specifically, the employ process includes making generalisations, constructing mathematical arguments, and

explaining or justifying mathematical results, while the interpret process involves evaluating whether mathematical conclusions are meaningful in context. This perspective confirms the importance of reasoning in mathematics assessment, but it also suggests that broad frameworks do not necessarily operationalise all reasoning indicators in a focused manner.

For this reason, more targeted assessment is needed to examine how students perform on particular aspects of reasoning and whether the instrument used to assess them functions appropriately. Prior studies have applied Rasch analysis to assess psychometric properties of tests in domains such as statistical reasoning (Ramadhani et al., 2022), statistical literacy (Riwayani et al., 2024), and creative thinking (Suherman & Vidákovich, 2022). Rasch modelling offers unique advantages by situating item difficulty and person ability on a common scale, detecting misfit items, and enabling analyses of bias through Differential Item Functioning (Aryadoust et al., 2021; Edwards & Alcock, 2010). Recent works have applied Rasch to higher-order thinking skills (Qirom & Nurlaelah, 2023; Suanto et al., 2023), reinforcing its potential to investigate reasoning in undergraduate contexts. In this study, numeracy-based tasks are used as the context for eliciting reasoning because they require students to interpret situations, identify mathematical relationships, and defend their conclusions. However, the main concern of this study is not numeracy as a separate construct, but undergraduate students' mathematical reasoning as reflected in those tasks. To address this concern, an assessment approach is needed that can evaluate not only students' scores, but also the functioning of each scored component and the fairness of the instrument across groups.

The Rasch model offers a suitable framework for this purpose because it places item difficulty and person ability on a common scale, evaluates the fit of each scored component, and enables the detection of potential bias through Differential Item Functioning analysis (Aryadoust et al., 2021; Edwards & Alcock, 2010). Compared with descriptive score analysis alone, Rasch modelling provides stronger evidence regarding whether an instrument measures the intended construct consistently and equitably. This is particularly important in the assessment of reasoning, where students may obtain similar total scores while demonstrating different patterns of strength and difficulty across conjecturing, generalising, and justifying.

Accordingly, this study investigates undergraduate students' mathematical reasoning in terms of conjecturing, generalising, and justifying through a Rasch model approach. The study aims to examine the functioning of the assessment instrument, identify the relative difficulty of the reasoning components, and determine whether the instrument operates fairly across gender, GPA, and semester level. The novelty of this study lies in combining a focused analysis of specific reasoning processes with Rasch-based evidence on item functioning and assessment fairness in a higher education context. On this basis, the study addresses the following research questions: (1) What does the distribution of item difficulty and student ability reveal about students' performance in conjecturing, generalizing, and justifying? And; (2) Does the instrument function equivalently across gender, GPA, and semester level?

Methods

Research design

This study employed a quantitative descriptive approach using the Rasch model to measure undergraduate students' mathematical reasoning skills. The Rasch model was selected for its ability to evaluate item-level performance and test fairness across demographic groups, aligning with the study's objectives.

Participants

The participants were 185 undergraduate students from the Mathematics Education program at a state university in Kediri, Indonesia. The sample included students from the 1st, 3rd, and 5th semesters, allowing comparisons across different levels of academic progression. All participants had completed prerequisite mathematics courses relevant to the reasoning tasks. The sampling technique is purposive. This selection process ensured that all participants had undergone analogous instructional experiences pertinent to their respective semesters. The demographic variables examined in the Rasch and DIF analyses were gender, cumulative grade point average (GPA), and semester level. To support the analysis more clearly, Table 1 should report the frequencies and percentages for all three variables used in the study.

Instruments

The instrument was developed by the researchers on the basis of the mathematical reasoning literature and was designed to assess three aspects of reasoning: conjecturing, generalising, and justifying. It consisted of two numeracy-based tasks. To ensure the validity, the instrument was validated by four lecturers specializing in mathematics education, each possessing a master's degree and averaging eight years of pedagogical experience within a university context. These experts reviewed the content and linguistic aspects of the instrument, ensuring its coherence with the fundamental elements of mathematical reasoning, specifically conjecturing, generalizing, and justifying. The validation procedure substantiated that the instrument effectively measured the intended constructs, thereby confirming its reliability for evaluating mathematical reasoning within the scope of this investigation.

Procedures

Data were collected during regular face-to-face classroom sessions. Students completed the tasks individually within a 100-minute time frame, which corresponds to a duration of two instructional hours, a temporal parameter deemed adequate for sustaining student focus and involvement throughout the evaluation. All responses were collected and anonymized before scoring. The responses were scored by researcher, using the analytics rubric. A scoring rubric was developed based on these dimensions, allocating a maximum of 15 points for conjecturing, 20 points for generalizing, and 20 points for justifying. The written responses were subsequently evaluated by the research team using a predefined scoring rubric. The scores were categorized according to the rubric, accessible through the provided link. To facilitate analysis, each aspect was coded as follows: C1, G1, and J1 (for Question 1) and C2, G2, and J2 (for Question 2). These codes represented the six reasoning components analyzed.

Data analysis

Descriptive statistical analysis was calculated to summarize the score distribution. Rasch modeling was performed using R software to evaluate item validity, individual ability, and test reliability. Item fit was assessed using several Rasch fit criteria, including Infit and Outfit Mean Square (MNSQ), Standardized Z-score (ZSTD), and Point Measure Correlation (PTMA). In accordance with established Rasch measurement guidelines, acceptable item fit is indicated by an MNSQ value between 0.5 and 1.5, a ZSTD value in the range of –2 to +2, and a positive PTMA value, reflecting each item's consistency with the overall construct (Boone, 2016; Bond et al., 2020). Wright maps were created to compare the distribution of item difficulty and student ability on a generalized logit scale.

The next step was a Differential Item Functioning (DIF) analysis to identify potential item bias across gender, GPA categories, and semester levels. DIF was evaluated using statistical significance criteria based on adjusted p-values, with items with p-values less than 0.05 considered to exhibit significant DIF. This analysis ensures that the instrument functions equivalently across subgroups and provides evidence of measurement invariance (Bond et al., 2020).

Results

Participant characteristics

The objective of this study was to standardized the mathematical reasoning skills test using the Rasch model. This standardized aimed to evaluate the feasibility of the test items in assessing students' ability. Previous studies have demonstrated the utility of Rasch analysis for both multiple-choice and essay-type items. In this study, Rasch analysis was applied to two essay questions designed to measure conjecturing, generalizing, and justifying. The characteristics of the respondents, disaggregated by gender, GPA category and Semester level are presented in Table 1.

Table 1. Participant characteristics by semester level

Semester	GPA Category	Score (Average)		n (count)	
		Male	Female	Male	Female
1	<3.25	-	28.42	-	7
	> 3.25	47.6	41.6	10	48
3	<3.25	-	40.34	-	3
	> 3.25	39.45	44.19	11	47
5	<3.25	19	-	1	-
	> 3.25	25	55.6	2	56

Table 1 showed the average student grades by semester level, GPA category, and gender. Overall, students with a GPA above 3.25 tended to achieve higher grades than students with lower GPAs across all semesters. The table also shows that there are several empty cells, such as first-semester students, students with a GPA less than 3.25, and male gender. These empty

cells indicated that in that category, no male students had a GPA below 3.25. The table also illustrates that the majority of female students' GPAs were higher than those of male students.

Item fit

The fit statistics of each test item are displayed in Table 2. All six items met the required infit and outfit MNSQ range of 0.5–1.5, confirming their validity. The infit values ranged from 0.73 (G1) to 1.49 (C1), while the outfit values ranged from 0.73 (J1) to 1.21 (C1). These results indicate that each item consistently measured the intended aspects of mathematical reasoning and was feasible for use in the test.

Table 2. Infit and outfit scores for the mathematical reasoning skill test

Item	Measure	Infit_MNSQ	Outfit_MNSQ	Infit_ZSTD	Outfit_ZSTD	PTMA
C1	-1.12	6.76	6.98	28.80	29.89	0.68
C2	0.41	12.96	7.19	59.82	30.96	0.60
G1	-2.09	4.97	2.30	19.87	6.51	0.63
G2	-0.55	3.85	1.93	14.26	4.67	0.63
J1	-0.13	4.85	2.37	19.25	6.87	0.58
J2	-0.09	3.79	2.17	13.96	5.82	0.58

The PTMA value has a positive value with a range of 0.577 to 0.676, indicating that each item has a positive correlation with the overall test score and contributes significantly to the measurement of mathematical reasoning skills. The item difficulty level has a range of -2.094 (easier items) to 0.411 logit (more difficult items), indicating a moderate level of distribution of item difficulty across the test.

However, the item fit statistics measured using the Infit MNSQ and Outfit MNSQ revealed values that still did not fit within the range in the Rasch analysis, which is between 0.5 and 1.5. Similarly, the ZSTD values were very high, far from the model. This result is likely due to inconsistent response patterns or poorly functioning rating scale categories. Therefore, although the PTMA values are acceptable, the overall item fit is not yet appropriate, and further refinement of the items such as revising the item wording or combining value categories is needed to improve the quality of the measurement

Distribution of item difficulty and student ability

The Wright map (Person-Item map) is used to simultaneously display the distribution of student ability and item difficulty on the same logit scale. Wright charts have an advantage over summary tables because they provide a visual comparison that allows researchers to examine the alignment between item difficulty and student ability levels. This is particularly useful for identifying whether test items are appropriately targeted to a sample.

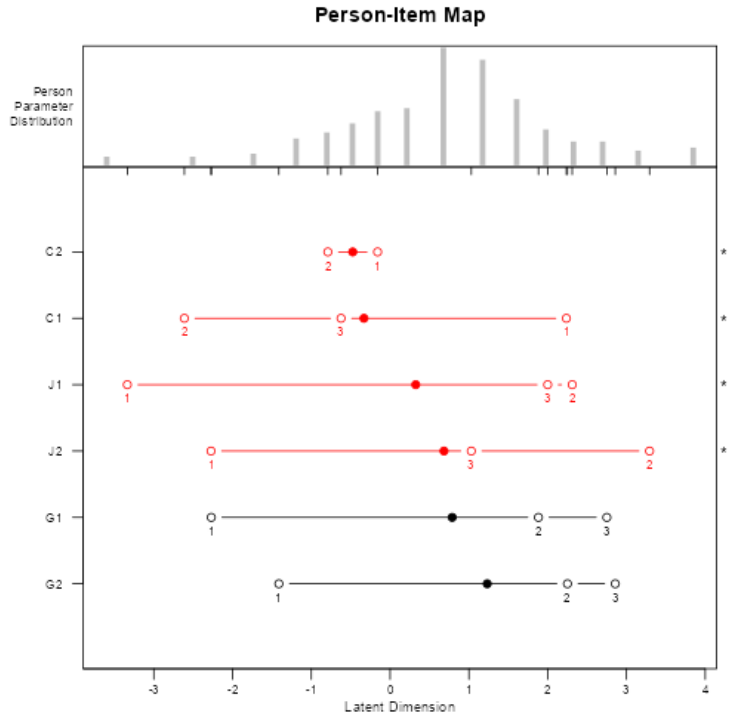

Figure 1. Person item map of mathematical reasoning skills test

Figure 1 provides a visual representation of the alignment between student ability and item difficulty on the same logit scale. The top panel shows the distribution of student ability, with most students clustered between –0.50 and +0.50 logit, indicating a moderate level of mathematical reasoning ability. The bottom panel displays item difficulty and category thresholds. Of the six items, J2 is positioned at the highest difficulty level (+1.25 logit), indicating that only students with relatively high ability are likely to answer this item correctly. In contrast, item C2 is positioned at –2.51 logit, indicating that it is very easy and can be answered correctly by almost all students. The map clearly shows a mismatch between student ability and item difficulty, particularly for item J2, which is above the ability range of most students. This suggests that tasks involving conjecture and justification are the most challenging aspects of mathematical reasoning for students. Furthermore, students with lower ability levels (below –1.00 logit) tended to be unsuccessful in answering questions of moderate difficulty, while only a small proportion of students with high ability were able to solve the most difficult questions.

Item characteristic curve (ICC)

The Item Characteristic Curves (ICC) in Figure 2 illustrate the relationship between student ability level and the probability of achieving a higher score on each item. All curves exhibit a generally increasing (monotonic) pattern, indicating that students with higher ability tend to have a greater probability of answering correctly. This suggests that, in general, the items functioned in the direction expected from the Rasch model. However, marked differences were observed in the slope and shape of the curves across items.

Some items exhibited relatively gentle slopes, indicating limited discrimination between students with different ability levels. This result is consistent with the item fit results, where high MNSQ Infit and Outfit values indicate poor fit to the Rasch model. Flatter curves indicate that students with similar ability levels may respond inconsistently, contributing to the improved fit statistics. Conversely, items with steeper curves demonstrated better discrimination, although they were still affected by overall poor fit. These findings imply that although the items have directional validity, their measurement precision is less than optimal, and revisions primarily in item clarity or scoring structure may be needed to improve the items' alignment with the Rasch model and enhance their ability to differentiate between different levels of mathematical reasoning.

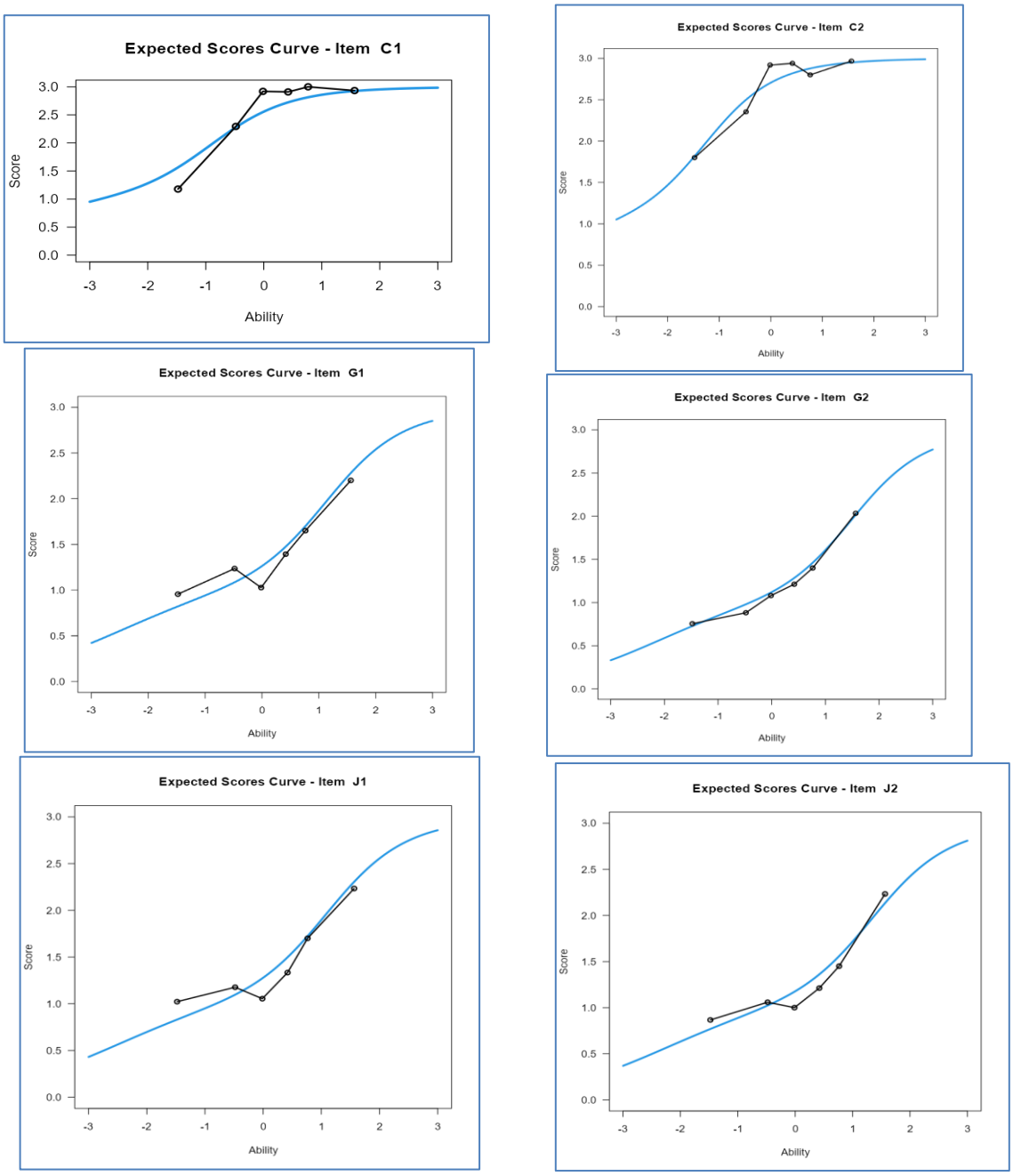

Figure 2. Item characteristic curve (ICC)

Differential item functioning (DIF)

Differential Item Functioning (DIF) analysis was conducted to examine whether items function differently across groups with similar ability levels. In this study, DIF was tested based on GPA and semester level. Items with significant differences (adjusted p-values <0.05) indicated potential bias. Table 3 show DIF analysis based on gender. All items produced p-values greater than 0.05 (range = 0.316–0.697), confirming that no item functioned differently for male and female students. This indicates that the instrument measured reasoning ability equivalently across gender groups.

Table 3. Differential item function test based on gender

Aspect	Statistic	p	Adj.p
C1	0.721	0.697	0.697
C2	1.502	0.472	0.697
G2	1.823	0.402	0.697
J2	0.766	0.682	0.697
G1	2.306	0.316	0.697
J1	1.447	0.485	0.697

Table 4 showed the results of the DIF analysis. DIF analysis based on semester level and GPA was used to determine whether the test items functioned fairly across groups with different academic backgrounds. Academic background, based on semester level, reflects students' learning experiences, while GPA and GPA represent academic performance. The analysis aimed to identify whether the test items were fair and able to measure mathematical reasoning consistently across groups, thus supporting their validity.

Table 4. Results of DIF analysis based on GPA and semester level

Group	Test Code	Statistic	Adj.p
GPA scale 3.01 – 3.5 and GPA > 3.5	C1	6.1688	0.275
	C2	2.5012	0.839
	G2	0.5219	0.924
	J2	1.1729	0.839
	G1	0.0116	0.994
	J1	1.1626	0.839
Semester level of 1st and 3rd	C1	3.995	0.319
	C2	3.673	0.319
	G2	1.041	0.713
	J2	0.110	0.946
	G1	4.549	0.319
	J1	1.556	0.689
Semester level of 1st and 5th	C1	0.969	0.861
	C2	0.728	0.861
	G2	7.426	0.146
	J2	0.299	0.861
	G1	2.373	0.861
	J1	0.504	0.861

Group	Test Code	Statistic	Adj.p
Semester level of 3rd and 5th	C1	1.449	0.681
	C2	6.353	0.250
	G2	3.737	0.463
	J2	0.853	0.681
	G1	0.769	0.681
	J1	2.009	0.681

The DIF analysis results in Table 4 indicate that none of the items exhibited significant differential functioning across the semester groups or GPA levels examined. All adjusted p-values were greater than 0.05, for both GPA-based and semester-based comparisons. These results indicate that items C1, C2, G1, G2, J1, and J2 functioned equivalently across student groups, meaning that students with similar ability levels had a similar probability of answering each item correctly, regardless of their GPA or semester level. Therefore, these test items can be considered fair and free from item bias when administered to students.

Discussion

Undergraduate students' mathematical reasoning ability in conjecturing, generalizing, and justifying

This study found that undergraduate students' mathematical reasoning skills remain uneven across the three aspects examined. Conjecturing and justifying emerged as the weakest areas, while generalizing was demonstrated only partially. These results confirm earlier studies that identified conjecturing and justification as persistent difficulties in reasoning development (Agustyaningrum et al., 2019; Paradesa, 2018). The findings suggest that while students can recognize patterns, they often fail to extend these into valid hypotheses or logically supported arguments.

Conjecturing requires learners to move beyond surface-level observations to form plausible general statements. In this study, many students were unable to produce initial hypotheses, echoing previous evidence that learners often struggle in the early stages of reasoning (Rohati et al., 2023). Even at the undergraduate level, where abstract thinking is expected, students remain dependent on empirical observations, consistent with Case and Speer (2021) findings that undergraduates perform better in applied contexts than in theoretical reasoning. This indicates a gap in instruction that needs explicit attention.

Regarding generalization, students demonstrated partial ability to identify patterns but struggled to broaden their reasoning across contexts. Similar results have been reported by Nurdiansah et al. (2024), who noted that students can verbalize observed patterns but rarely extend them into formal mathematical structures. Subanji et al. (2021) emphasize that numeracy-based problems can scaffold generalization, but only when students are supported to connect instances into broader principles. Embedding academic numeracy frameworks across courses (Adelia et al., 2024; Galligan, 2013) could systematically enhance this dimension of reasoning.

Justification was the most problematic area. Many prospective teachers showed a limited ability to go beyond their initial conjectures, struggling to build logically sound arguments based on formal mathematical principles (Hamidah, Susiswo, H, et al., 2025). Such reliance on empirical reasoning has been observed widely among teacher candidates (Brečka et al., 2022; Rosyadi et al., 2022). This suggests that even prospective educators may not yet value the process of verification as much as the final result. Instruction that explicitly frames justification as a form of critical evaluation, such as the approach proposed by Jain and Rogers (2019) and discursive multimodal methods integrating language with mathematical meaning (Clarke, 2024), could help improve justification skills.

Taken together, these findings indicate that undergraduate students require targeted instructional strategies to strengthen reasoning. Ellis et al. (2022) argue that learning designs should scaffold the transition from empirical reasoning to structural proof. Without such scaffolds, students risk plateauing at surface-level reasoning, which limits their capacity to engage in higher-order problem solving and to function as future mathematics educators. Furthermore, incorporating real-world problem-solving situations into the curriculum may boost engagement and practical comprehension, equipping educators to meet the evolving demands of modern classrooms (Hamidah, Susiswo, Susanto, & Osman, 2025).

Mathematical reasoning of undergraduate students based on gender and GPA

Although the Rasch model confirmed that the assessment instrument was unbiased, qualitative analysis revealed distinct reasoning tendencies between male and female students. Female students displayed higher accuracy and stronger adherence to structured strategies, consistent with prior studies noting systematic approaches among women (Rokhima et al., 2019; Rubianti et al., 2022). Male students, meanwhile, exhibited flexibility in trying alternative strategies, but often with reduced precision. This reflects earlier findings that gender-based cognitive orientations influence reasoning styles (Rusdi et al., 2020; Tariq et al., 2013).

These gendered differences align with broader psychological and sociocultural explanations. Research suggests that affective factors, such as math anxiety and neuroticism, may partly explain variations in numeracy performance between male and female students (Lunardon et al., 2022; Vos et al., 2023). At the same time, gender stereotypes in mathematics continue to shape students' self-concepts and motivation (Li et al., 2022), potentially reinforcing these patterns. The convergence of these factors highlights the complexity of gender in reasoning performance.

GPA was found to be positively associated with reasoning performance, as students with higher academic achievement showed greater logical coherence and problem-solving organisation. This finding is in line with earlier evidence linking reasoning abilities to overall academic success (Darta & Saputra, 2018; Hasanah et al., 2019). Luo et al. (2021) further demonstrated that reasoning ability significantly predicts achievement in mathematics and science, suggesting a reciprocal relationship between performance and reasoning development.

Nevertheless, GPA alone cannot fully explain reasoning differences. Recent studies suggest that non-cognitive factors such as conscientiousness, self-efficacy, and motivation play

mediating roles (Carvalho, 2016; Minnigh & Coyle, 2023). Students with lower GPAs may therefore benefit from scaffolding that not only addresses content knowledge but also fosters metacognitive and affective skills. Such approaches could help reduce disparities and ensure that reasoning abilities are developed equitably.

The implications of these findings are significant for pedagogy. Gender-sensitive instructional practices can be designed to enhance precision among male students while encouraging flexibility among female students, ensuring that both cognitive orientations are valued (Rubianti et al., 2022; Wrigley-Asante et al., 2023). Similarly, students with lower GPAs may require additional support through structured problem-solving frameworks, collaborative learning, and reflective practice. By acknowledging these demographic factors, educators can implement more nuanced teaching strategies.

Finally, this study illustrates the methodological value of Rasch analysis in assessing reasoning fairly. By confirming item validity and reliability, identifying item difficulty through Wright maps, and ruling out demographic bias via DIF, Rasch modelling provides a rigorous framework for evaluating reasoning in higher education (Qirom & Nurlaelah, 2023; Ramadhani et al., 2022; Suanto et al., 2023). These results extend previous Rasch applications beyond psychometric validation, offering a means to classify reasoning hierarchies and inform equitable instructional design.

In conclusion, while this study confirms the fairness of the assessment tool, it also reveals meaningful differences in reasoning performance across conjecturing, generalizing, and justifying, as well as gender- and GPA-related tendencies. These findings underscore the importance of targeted instructional strategies and curriculum designs that foster higher-order reasoning skills in undergraduate mathematics education.

Conclusion

This study revealed that undergraduate students faced the greatest challenges in conjecturing and justifying, while their ability to generalize was only partially developed. The Rasch analysis confirmed the validity, reliability, and fairness of the instrument across gender, GPA, and semester levels. Nevertheless, observed tendencies showed that female students tended to be more accurate and structured, male students more flexible but less precise, and students with higher GPAs demonstrated more coherent reasoning. These findings imply the need for curricula and instructional strategies that explicitly foster higher-order reasoning and promote equitable development of conjecturing, generalizing, and justifying skills.

This research is limited by its focus on a single institution, the use of only two numeracy-based tasks, and the absence of longitudinal tracking. Future studies should expand the sample across multiple universities, employ a broader range of reasoning tasks, and adopt longitudinal designs to capture the progression of reasoning over time. Further investigations may also integrate psychological and contextual factors to better understand how demographic and non-cognitive variables influence reasoning. The contribution of this study lies in providing evidence for the fair measurement of reasoning skills and offering insights that can inform curriculum design and pedagogy in mathematics education.

Acknowledgment

The authors extend their sincere appreciation to the prospective teachers who willingly participated as respondents in this study, thereby contributing significantly to the smooth and effective implementation of the research.

Declarations

Conflicts of Interest	:	The authors declare no conflict of interest.
Generative AI Statement	:	AI Used for Limited, Non-Substantive Support: Generative AI tools, such as Grammarly was employed solely for language editing and minor phrasing enhancements. Besides, conceptualization and analysis were independently developed and verified by the authors. Some scholarly content found by Scopus AI.
Funding Statement	:	This work received no specific grant from any public, commercial, or not-for-profit funding agency.
Author Contributions	:	**Dewi Hamidah:** Conceptualization, Data Curation, Investigation, Formal Analysis, Methodology, Writing–Original Draft Preparation, and Writing–Editing; **Jerhi Wahyu Fernanda:** Methodology, Formal Analysis, Visualization, Writing–Review & Editing, and Validation; **Zun Azizul Hakim:** Methodology, and Writing–Review & Editing, and Formal Analysis; **Galuh Nuril Latifah:** Methodology and Validation.

References

Adelia, V., Putri, R. I. I., & Zulkardi. (2024). A systematic literature review: How do we support students to become numerate? *International Journal of Evaluation and Research in Education*, *13*(2), 842–851. https://doi.org/10.11591/ijere.v13i2.26299

Agustyaningrum, N., Kusmayadi, T. A., & Riyadi. (2019). Student's conjecturing ability in mathematical problem solving. *Journal of Physics: Conference Series*, *1188*(1), 012031. https://doi.org/10.1088/1742-6596/1188/1/012031

Andrews-Larson, C., Johnson, E., Peterson, V., & Keller, R. (2021). Doing math with mathematicians to support pedagogical reasoning about inquiry-oriented instruction. *Journal of Mathematics Teacher Education*, *24*(2), 127–154. https://doi.org/10.1007/s10857-019-09450-3

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Brečka, P., Valentová, M., & Lančarič, D. (2022). The implementation of critical thinking development strategies into technology education: The evidence from Slovakia. *Teaching and Teacher Education*, *109*, 103555. https://doi.org/10.1016/j.tate.2021.103555

Carvalho, R. G. G. (2016). Gender differences in academic achievement: The mediating role of personality. *Personality and Individual Differences*, *94*, 54–58. https://doi.org/10.1016/j.paid.2016.01.011

Case, J., & Speer, N. (2021). Calculus students' deductive reasoning and strategies when working with abstract propositions and calculus theorems. *PRIMUS*, *31*(2), 184–201. https://doi.org/10.1080/10511970.2019.1660931

Clarke, J. (2024). Mathematical meaning construction in language, tables, and images: The potential utility of language proficiency development in numeracy teaching. *Critical Studies in Teaching and Learning*, *12*(1), 56–74. https://doi.org/10.14426/cristal.v12i1.800

Darta, D., & Saputra, J. (2018). Indicators that influence prospective mathematics teachers' representational and reasoning abilities. *Journal of Physics: Conference Series*, *948*(1), 012053. https://doi.org/10.1088/1742-6596/948/1/012053

Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its Applications*, *29*(4), 165–175. https://doi.org/10.1093/teamat/hrq008

Ellis, A., Özgür, Z., & Reiten, L. (2022). From empirical to structural reasoning: Shifts in students' justifications in mathematics. *Journal of Mathematical Behavior*, *66*, 100940. https://doi.org/10.1016/j.jmathb.2021.100940

Galligan, L. (2013). A systematic approach to embedding academic numeracy at university. *Higher Education Research and Development*, *32*(5), 734–747. https://doi.org/10.1080/07294360.2013.777038

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM Mathematics Education*, *47*(4), 531–548. https://doi.org/10.1007/s11858-015-0708-1

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

Hamidah, D., Susiswo, S., Susanto, H., & Osman, S. (2025). Prospective mathematics teachers' mathematization competencies in solving word problems. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, *9*(3), 894–908. https://doi.org/10.22437/jiituj.v9i3.33211

Hasanah, S. I., Tafrilyanto, C. F., & Aini, Y. (2019). Mathematical reasoning: The characteristics of students' mathematical abilities in problem solving. *Journal of Physics: Conference Series*, *1188*(1), 012057. https://doi.org/10.1088/1742-6596/1188/1/012057

Herbert, S., & Williams, G. (2023). Eliciting mathematical reasoning during early primary problem solving. *Mathematics Education Research Journal*, *35*(1), 77–103. https://doi.org/10.1007/s13394-021-00376-9

Jain, P., & Rogers, M. (2019). Numeracy as critical thinking. *Adults Learning Mathematics: An International Journal*, *14*(1), 20–31. https://www.alm-online.net/images/ALM/journals/almij-volume14-1-june2019.pdf

Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics*, *96*(1), 1–16. https://doi.org/10.1007/s10649-017-9761-8

Joachin-Arizmendi, I., Espinoza, E. L., Carballo, A. M., & Reyna-Hernández, G. (2024). Diagnostic study of mathematical reasoning in novice university students. *International*

Electronic Journal of Mathematics Education, *19*(3), 0788. https://doi.org/10.29333/iejme/14862

Kadarisma, G., Nurjaman, A., Sari, I. P., & Amelia, R. (2019). Gender and mathematical reasoning ability. *Journal of Physics: Conference Series*, *1157*(4), 042109. https://doi.org/10.1088/1742-6596/1157/4/042109

Li, J., Faisal, E., & Al Hariri, A. (2022). Numbers for boys and words for girls? Academic gender stereotypes among Chinese parents. *Sex Roles*, *87*(9–10), 475–489. https://doi.org/10.1007/s11199-022-01298-0

Lunardon, M., Cerni, T., & Rumiati, R. I. (2022). Numeracy gender gap in STEM higher education: The role of neuroticism. *Frontiers in Psychology*, *13*, 856405. https://doi.org/10.3389/fpsyg.2022.85640

Luo, M., Sun, D., Zhu, L., & Yang, Y. (2021). Evaluating scientific reasoning ability: Student performance and the interaction effects between grade level, gender, and academic achievement level. *Thinking Skills and Creativity*, *40*, 100820. https://doi.org/10.1016/j.tsc.2021.100820

Minnigh, T. L., & Coyle, T. R. (2023). Gender differences in self-efficacy partially explain the female underprediction effect. *Journal of Research in Personality*, *104*, 104214. https://doi.org/10.1016/j.jrp.2023.104214

Nurdiansah, I., Rahardjo, S., & Susiswo. (2024). Exploring students' mathematical generalization ability in problem solving. *Journal of Physics: Conference Series*, *2600*(1), 012035. https://doi.org/10.1088/1742-6596/2600/1/012035

OECD. (2022). PISA 2022 Mathematics Framework. https://pisa2022-maths.oecd.org/

Paradesa, R. (2018). Students' ability in generalization of mathematical pattern. *Infinity Journal*, *7*(1), 21–30. https://doi.org/10.22460/infinity.v7i1.p21-30

Purnomo, H., Sa'dijah, C., Hidayanto, E., Sisworo, P., H, & Anwar, L. (2022). Development of instrument numeracy skills test of minimum competency assessment (MCA) in Indonesia. *International Journal of Instruction*, *15*(3), 635–648. https://doi.org/10.29333/iji.2022.15335a

Qirom, M., & Nurlaelah, E. (2023). Rasch model: Analysing the items quality of mathematics higher-order thinking skill instrument. *Jurnal Pendidikan dan Pembelajaran*, *30*(1), 11–20. https://doi.org/10.17977/um047v30i12023p011

Ramadhani, R., Saragih, S., & Napitupulu, E. E. (2022). Exploration of students' statistical reasoning ability in the context of ethnomathematics: A study of the Rasch model. *Mathematics Teaching Research Journal*, *14*(1), 138–168.

Riwayani, Istiyono, E., Supahar, Perdana, R., & Soeharto. (2024). Analyzing students' statistical literacy skills based on gender, grade, and educational field. *International Journal of Evaluation and Research in Education*, *13*(2), 842–851. https://doi.org/https://doi.org/10.11591/ijere.v13i2.26299

Rohati, R., Kusumah, Y. S., & Kusnandi, K. (2023). Exploring students' mathematical reasoning behavior in junior high schools: A grounded theory. *Education Sciences*, *13*(3), 252. https://doi.org/10.3390/educsci13030252

Rokhima, W. A., Kusmayadi, T. A., & Fitriana, L. (2019). Mathematical reasoning of student in senior high school based on gender differences. *Journal of Physics: Conference Series*, *1318*(1), 012092. https://doi.org/10.1088/1742-6596/1318/1/012092

Rosyadi, A. A. P., Sa'dijah, C., Susiswo, S., & Rahardjo, S. (2022). Critical thinking of prospective mathematics teachers: What are the errors in argumentation? *Multicultural Education*, *8*(3), 80–90. https://doi.org/10.5281/zenodo.7348010

Rubianti, N. S., Usodo, B., & Subanti, S. (2022). The difference of mathematical reasoning between male and female students. *AIP Conference Proceedings*, *2566*(1), 030005. https://doi.org/10.1063/5.0117121

Rusdi, M., Fitaloka, O., Basuki, F. R., & Anwar, K. (2020). Mathematical communication skills based on cognitive styles and gender. *International Journal of Evaluation and Research in Education*, *9*(4), 847–856. https://doi.org/10.11591/ijere.v9i4.20497

Suanto, H., Mulyana, T., & Prasetyo, H. (2023, 2023). *Analysis of students' mathematical reasoning instrument using Rasch model* Proceedings of the International Conference on Education, Social Sciences and Humanities (ICoESSE 2023, https://doi.org/10.2991/978-2-494069-55-8_18

Subanji, N., T, R., D, & Purnomo, H. (2021). The statistical creative framework in descriptive statistics activities. *International Journal of Instruction*, *14*(2), 591–608. https://doi.org/10.29333/iji.2021.14233a

Suherman, S., & Vidákovich, T. (2022). Tapis patterns in the context of ethnomathematics to assess students' creative thinking in mathematics: A Rasch measurement. *Mathematics Teaching Research Journal*, *14*(4), 56–79.

Tariq, V. N., Qualter, P., Roberts, S., & Barnes, L. (2013). Mathematical literacy in undergraduates: Role of gender, emotional intelligence, and emotional self-efficacy. *International Journal of Mathematical Education in Science and Technology*, *44*(7), 1010–1029. https://doi.org/10.1080/0020739X.2013.783632

Vos, H., Marinova, M., Léon, S. C., Sasanguie, D., & Reynvoet, B. (2023). Gender differences in young adults' mathematical performance: Examining the contribution of working memory, math anxiety, and gender-related stereotypes. *Learning and Individual Differences*, *102*, 102255. https://doi.org/10.1016/j.lindif.2022.102255

Wrigley-Asante, C., Ackah, C. G., & Frimpong, L. K. (2023). Gender differences in academic performance of students studying STEM subjects at the University of Ghana. *SN Social Sciences*, *3*(6), 91. https://doi.org/10.1007/s43545-023-00834-5