Prima: Jurnal Pendidikan Matematika
Vol. 10, No. 3, September 2026, pp. 770 - 783
P-ISSN: 2579-9827, E-ISSN: 2580-2216
Web: http://jurnal.umt.ac.id/index.php/prima

RASCH MODEL ANALYSIS OF MATHEMATICAL REASONING ABILITY ITEMS ON LINEAR EQUATIONS FOR JUNIOR HIGH SCHOOL STUDENTS USING MINISTEP

Widuri Ambaika[1], Desy Lusiyana[2], Rifqi Hidayat[3]
[123]Universitas Muhammadiyah Cirebon, Jl.Fatahillah No.40 , West Java, Indonesia
e-mail: ambaikawiduri@gmail.com

Abstract

The limited availability of valid and reliable assessment instruments for measuring junior high school students' mathematical reasoning ability on linear equations motivated this study. Despite widespread recognition of mathematical reasoning as a core competency in school mathematics, few instruments in this domain have been rigorously validated using modern measurement approaches such as the Rasch Model. This study aimed to analyze the item quality of a mathematical reasoning test on linear equations using the Rasch Model via Ministep software, focusing on item validity, instrument reliability, difficulty index, and person-item mapping. A descriptive quantitative approach was conducted with 30 junior high school students chosen through purposive sampling. Four polytomous essay items were developed based on NCTM's four reasoning indicators, namely drawing conclusions, proposing conjectures, constructing proofs or arguments, and explaining solution steps. Results showed all four items valid: Outfit MNSQ (0.71–0.97), Outfit ZSTD (−0.34 to 0.15), and Point Measure Correlation (0.82–0.84). Person Reliability was 0.87 (separation index 2.63), classified as good. All items were of moderate difficulty (logit: 0.00 to +0.70). The Wright Map indicated students' average ability exceeded item difficulty, suggesting the need for harder items in future development. Based on the Rasch analysis, the instrument demonstrates acceptable measurement properties and is deemed appropriate for formative assessment of mathematical reasoning in linear equation topics, while also pointing to clear directions for future instrument refinement.

Keywords: mathematical reasoning ability, linear equation, rasch model, ministep

Abstrak

Penelitian ini dilatarbelakangi oleh masih terbatasnya ketersediaan instrumen penilaian yang valid dan reliabel untuk mengukur kemampuan penalaran matematis siswa SMP pada materi persamaan garis lurus. Meskipun kemampuan penalaran matematis diakui sebagai kompetensi inti dalam pembelajaran matematika sekolah, masih sedikit instrumen pada topik ini yang tervalidasi secara ketat menggunakan pendekatan pengukuran modern seperti Model Rasch. Penelitian ini bertujuan untuk menganalisis kualitas butir soal tes penalaran matematis pada materi persamaan garis lurus menggunakan Model Rasch melalui perangkat lunak Ministep, dengan fokus pada validitas butir, reliabilitas instrumen, indeks kesukaran, dan pemetaan person-item. Pendekatan kuantitatif deskriptif digunakan dengan melibatkan 30 siswa SMP yang dipilih melalui purposive sampling. Empat butir soal esai politomus dikembangkan berdasarkan empat indikator penalaran NCTM, yaitu menarik kesimpulan, mengajukan dugaan, menyusun bukti atau argumen, dan menjelaskan langkah penyelesaian. Hasil penelitian menunjukkan keempat butir soal valid, memenuhi kriteria Outfit MNSQ (0,71–0,97), Outfit ZSTD (−0,34 sampai 0,15), dan Point Measure Correlation (0,82–0,84). Reliabilitas person sebesar 0,87 dengan indeks separasi 2,63, tergolong baik. Seluruh butir berada pada tingkat kesukaran sedang (logit: −0,00 sampai +0,40). Wright Map menunjukkan rata-rata kemampuan siswa melebihi tingkat kesukaran butir, sehingga diperlukan pengembangan butir yang lebih sulit pada penelitian selanjutnya. Berdasarkan analisis Rasch, instrumen ini terbukti memiliki sifat pengukuran yang baik dan layak digunakan untuk asesmen formatif kemampuan penalaran matematis pada materi persamaan garis lurus, sekaligus memberi arah bagi penyempurnaan instrumen di masa mendatang.

Kata kunci: kemampuan penalaran matematis, persamaan garis lurus, model rasch, ministep

Received: July 15, 2026; Revised: September 19, 2026; Accepted: October 21, 2026

INTRODUCTION

Mathematical reasoning ability is a core competency in junior high school mathematics, as emphasized by the National Council of Teachers of Mathematics (NCTM, 2000). This ability includes: (1) drawing conclusions based on mathematical facts, (2) proposing conjectures, (3) constructing mathematical arguments, and (4) explaining solution steps systematically. These competencies are essential for students to understand algebraic structures and solve non-routine problems (Mukuka et al., 2023).

The topic of linear equations, situated within the algebra domain, demands integrated use of reasoning across symbolic, graphical, and numerical representations(Dwi Ikasari & Firmansyah, 2023). Students must analyze variable relationships, interpret slope and intercept, and apply logical inference (Asoraya & Ruli, 2023). However, multiple studies report persistently low achievement in algebraic reasoning among junior high school students, with weaknesses particularly evident in tasks requiring justification and proof (Aprilia et al., 2023).

A critical factor underlying these gaps is the absence of psychometrically validated instruments specifically designed to assess mathematical reasoning on linear equations (Sumintono, 2015). Instruments lacking rigorous validation can produce biased scores that misrepresent students' actual ability levels, leading to flawed instructional decisions (Sajdah & Hidayati, 2026). The Rasch Model addresses this limitation by providing sample-independent, objective measurement with simultaneous calibration of item validity, reliability, and difficulty (Siregar et al., 2025).

Recent applications of Rasch analysis in mathematics education have demonstrated strong outcomes. (Hidayat et al., 2025) achieved a reliability of 0.80 for a Rasch-based algebraic numeracy instrument, while (Rusyid et al., 2024). The findings confirmed that Rasch modelling effectively evaluated assesment items measuring students algebraic reasoning at the junior high school level.(Wardani et al., 2025) and (Ambarwati & Maarif, 2025) similarly validated numeracy instruments using Rasch with high separation indices. (Qirom & Nurlaelah, 2023) applied Rasch modelling to analyze HOTS-based mathematics items, while (Sarifah et al., 2024) demonstrated Rasch's utility in cognitive test validation at the elementary level.

However, a gap remains: no prior study has applied Rasch modeling specifically to mathematical reasoning on linear equations at the junior high school level. The current study

contributes to existing literature by constructing and validating an instrument, thereby complementing the work of (Hidayat et al., 2025) and extending its scope to the reasoning domain. The novelty lies in the integration of NCTM's four reasoning indicators with polytomous Rasch calibration for this specific mathematical topic.

The present research investigates the measurement properties of mathematical reasoning assesment instruments on linear equations for junior high school students using the Rasch Model with Ministep software, focusing on item validity, reliability, difficulty index, and person-item mapping.

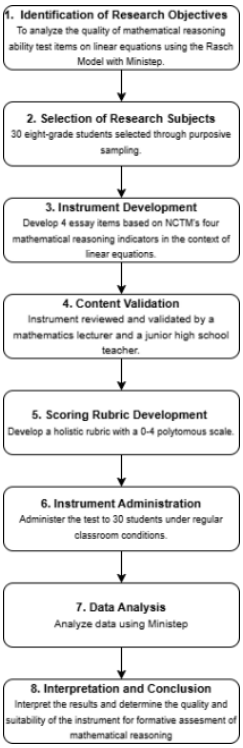
METHODS

This research applied a quantitative descriptive framework to investigate the measurement properties of a mathematical reasoning test through Rasch Model (McCombes, 2023). The quantitative approach was selected because Rasch modeling requires statistical calibration of item and person parameters, while the descriptive design serves to characterize instrument quality rather than test directional hypotheses.

Thirty junior high school students (Grade 9) were selected via purposive sampling from a single school in Cirebon, Indonesia, based on criteria of having completed the linear equations unit and representing a range of ability levels. The instrument comprised four essay items, each targeting one of the four (NCTM, 2000)mathematical reasoning indicators: (1) drawing conclusions from mathematical facts, (2) proposing conjectures, (3) constructing mathematical proofs or arguments, and (4) explaining solution steps sequentially. Items were embedded in linear equation contexts. The use of four essay items was based on the study's initial instrument validation design, in which each item was deliberately aligned with one of the four NCTM mathematical reasoning indicators. This one-to-one alignment was intended to ensure direct representation of each targeted reasoning component while maintaining a manageable assessment format for the preliminary Rasch calibration. Although four items are sufficient to provide an initial empirical examination of item functioning across the four indicators, this limited item pool does not comprehensively represent the full range of mathematical reasoning ability. Therefore, the present instrument should be regarded as an initial measurement model rather than a fully developed standardized assessment. Further development should include additional items for each indicator across different cognitive demands and difficulty levels to improve measurement coverage and diagnostic precision.

Expert validation was conducted by one mathematics education lecturer and one experienced mathematics teacher at the junior high school level to evaluate readability, indicator suitability, and contextual relevance (Wiyarsi et al., 2020).

Students responses were evaluated using a hilistic scoring rubric based on a polytomous



scale ranging from 0-4, structured hierarchically: 0 (no response or completely incorrect), 1 (shows minimal reasoning), 2 (partial reasoning with errors), 3 (mostly correct reasoning with minor gaps), and 4 (complete and correct reasoning). This scale yields richer response differentiation than dichotomous scoring, improving parameter estimation precision (Peeters & Augustine, 2023).

The implementation of the study followed an eight stage procedure: Identification of Research Objectives, Selection of Research Subjects, Instrument Development, Content Validation, Scoring Rubric Development, Instrument Administration, Data Analysis, Interpretation and Conclusion.

Table 1. Reasoning Ability Instrument

No.	Aspect	Indicator	Cognitive Level	Items
1	Inductive Reasoning	Drawing Conclusions	C4 (Analyzing)	A straight line passes through the points (0,2) and (2,6). In addition, there are several other lines with the

Figure 1. Research Procedure

				same gradient as this line but with different y-intercepts, namely: y=2x+1, y=2x+3 and y=2x+5. Plot these lines on the Cartesian plane, then observe their positions. Based on your observations, draw a conclusion whether the lines are parallel or intersecting.
2	Inductive Reasoning	Proposing Conjectures	C5 (Evaluating)	Given several straight lines, each defined by a pair of points lying on it. Line g_1 passes through points (0,1) and (2,5). Line g_2 passes through points (0,2) and (2,6). Line g_3 passes through points (0,3) and (2,7). Based on the pattern formed, determine the equations of lines g_1, g_2, and g_3 using more than one method.
3	Deductive Reasoning	Constructing Proofs or Arguments	C5 (Evaluating)	Consider the following two linear equations: $y = -3x + 6$ $3x + y - 6 = 0$ Prove that these equations represent the same straight line.
4	Deductive Reasoning	Explaining Solution Steps Coherently and Logically	C4 (Analyzing)	Given the linear equation $-3 = 2(x - 1)$. Convert this equation into the form $y = mx + c$ then explain the solution steps clearly.

RESULTS AND DISCUSSION

1. Results

Rasch analysis of the four-item mathematical reasoning instrument was conducted with 30 junior high school respondents using Ministep 4.8.2.0. Results are organized around three benchmarks: item validity, instrument reliability, and item difficulty, followed by Wright Map interpretation and discussion against prior literature.

1.1.Item Validity

Item validity was assessed against the three Rasch criteria presented in Table 2.

Table 2. Rasch Model Criteria For Item Quality

No.	Criteria	Value
1	Outfit MNSQ	0.5 < Outfit MNSQ < 1.5
2	Outfit ZSTD	-2.0 < Outfit ZSTD < +2.0
3	Pt Measure Corr.	0.4 < Pt Measure Corr < 0.85

Table 3. Item Validity

Figure 2. Results of Item Validity Analysis Using the Rasch Model

Item	Outfit MNSQ Value	Outfit ZSTD Value	Outfit Pt Measure Correlation Value	Category
1b	0.89	0.02	0.82	Valid
2b	0.77	-0.20	0.84	Valid
3b	0.97	0.15	0.83	Valid
4b	0.71	-0.34	0.83	Valid

Table 2 presents the full Item Statistics output from Ministep. All four items satisfied all three Rasch criteria, confirming their validity. Outfit MNSQ values (0.71–0.97) indicate

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
3	67	30	.40	.43	1.09	.42	.97	.15	.83	.83	65.5	74.5	3b
2	69	30	.04	.42	.89	-.44	.77	-.20	.84	.83	79.3	70.7	2b
1	70	30	-.13	.42	1.12	.60	.89	.02	.82	.83	86.2	70.3	1b
4	71	30	-.31	.41	.84	-.70	.71	-.34	.83	.83	69.0	70.6	4b
MEAN	69.3	30.0	.00	.42	.98	.0	.84	-.1			75.0	71.5	
P.SD	1.5	.0	.26	.01	.12	.6	.10	.2			8.2	1.7	

productive responses without overfit or underfit. Outfit ZSTD values (−0.34 to 0.15) fall well within the ±2.0 boundary. Point Measure Correlations (0.82–0.84) confirm strong item-ability alignment, consistent with findings by (Rusyid et al., 2024) who report convergent validity through Rasch fit statistic, and (Darmana et al., 2023) who similarly validated science instruments using the Rasch Model. Overall, these results confirm that all four items possess

strong construct validity and are free from response inconsistencies. This indicates that each item consistently measures the intended construct of mathematical reasoning ability, thereby providing sufficient empirical support for the instrument to be used in assessing students' reasoning on linear equations.

1.2. Instrument Realibility Analysis

Table 4 presents the reliability classification criteria used in this study.

Table 4. Reliability of Persons and Items

Reliability of Persons and Items	Criteria
>0,94	Excellent
0,91 – 0,94	Very Good
0,81 – 0,90	Good
0,67 – 0,80	Fair
>0,67	Poor

Person Reliability was 0.87 and Cronbach's Alpha (KR-20) was 0.87, both classified as good per Table 3. These values exceed the 0.80 threshold reported by (Hidayat et al., 2025) for an algebraic numeracy instrument and are consistent with (Ambarwati & Maarif, 2025) who also report comparable reliability values in numeracy instrument. The separation measure of 2.63 reflects the instruments ability to discriminate among respondents across at least two separate levels of ability. The polytomous 0–4 scoring scale contributed to higher reliability by capturing finer distinctions in reasoning quality compared to dichotomous formats (Peeters & Augustine, 2023). These results confirm that the instrument possesses

Figure 3. Instrument Realibility Summary

good reliability, both in terms of person and item consistency. This indicates that the

```
     SUMMARY OF 29 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       9.4       4.0        -.50    1.49       .83   -.22    .84   -.22 |
|  SEM        .5        .0         .96     .20       .13    .23    .13    .23 |
| P.SD       2.7        .0        5.08    1.08       .68   1.22    .68   1.23 |
| S.SD       2.7        .0        5.17    1.10       .69   1.24    .69   1.25 |
| MAX.      14.0       4.0        6.16    4.18      2.84   3.66   2.76   3.59 |
| MIN.       5.0       4.0       -9.13     .98       .00  -1.97    .00  -1.97 |
|-----------------------------------------------------------------------------|
| REAL RMSE   1.88 TRUE SD    4.72  SEPARATION  2.50  Person RELIABILITY  .86 |
|MODEL RMSE   1.84 TRUE SD    4.74  SEPARATION  2.57  Person RELIABILITY  .87 |
| S.E. OF Person MEAN = .96                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .87  SEM = 1.00
STANDARDIZED (50 ITEM) RELIABILITY = .99
```

instrument can be relied upon to produce stable and consistent measurement results when used to assess students' mathematical reasoning ability across different testing occasions.

1.3. Item Difficulty Index

The difficulty index was determined based on each item's logit value, following the categorization proposed by (Istiqomah et al., 2025), in which items with a logit value greater than +1 are classified as difficult, items ranging from −1 to +1 are classified as moderate, and items below −1 are classified as easy.

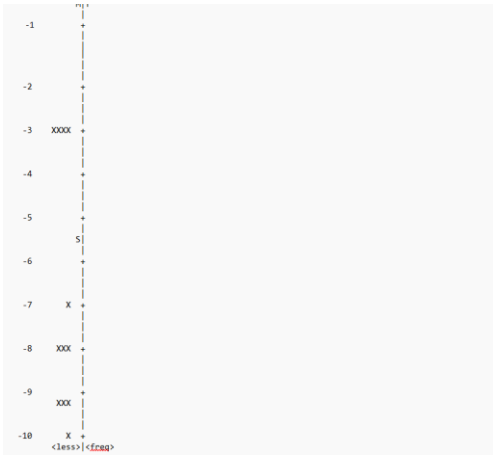
Figure 4 presents a Wright Map that depicts both participant ability and item difficulty within the same logic metric. The left section represents the distribution of respondents abilities, whereas the right section displays the relative difficulty of the items.

The Wright Map revealed that the instrument adequately measured student ability within the average range. The four items (1b, 2b, 3b, 4b) were clustered within 0.00 to +0.70 logits, corresponding to the moderate difficulty band only. A substantial number of respondents with higher ability (above +1.00 logit) were not matched by any item of appropriate difficulty, while respondents at the lower end also lacked sufficiently easy items.

These findings suggest the instrument requires further development, particularly



Figure 4. Item Difficult Index



through adding items with higher difficulty (logit > +1) to challenge high-ability students, and

items with lower difficulty (logit < −1) to better capture lower-ability performance. This recommendation aligns with (Wardani et al., 2025) and (Ambarwati & Maarif, 2025), who advocate for expanding item difficulty ranges in Rasch-based instruments. The limited item pool and sample size also constrained the stability of person ability estimates (Bond & Fox, 2013).

Table 5. Summary of Item Difficulty Index

Item No	Criteria	Value
1b	+0.20	Moderate
2b	+0.30	Moderate
3b	+0.70	Moderate
4b	0.00	Moderate

The item difficulty ranged from 0.00 logits (Item 4b) to +0.70 logits (Item 3b), covering a narrow range of 0.70 logits. Item 3b was the most difficult with a logit value of +0.70, positioned above the item mean, indicating that this item demands the highest level of reasoning ability and is designed to measure the indicator of constructing mathematical arguments. Items 2b and 1b were categorized as moderate, with logit values of +0.30 and +0.20 respectively, suitable for measuring student ability around the average level. Item 4b was the easiest, with a logit value of 0.00, aligned with the item mean calibration reference point, measuring the indicator of explaining solution steps sequentially. All four items fall within the moderate category (−1 ≤ logit ≤ +1), with none classified as difficult (logit > +1) or easy (logit < −1), consistent with (Rusyid et al., 2024) and (Qirom & Nurlaelah, 2023).

The concentration of all items within the narrow logit range of 0.00 to +0.70 indicates that the instrument was predominantly designed to measure ability at the moderate level. The person mean ability was slightly above the item mean (0.00 logit), indicating that the instrument was marginally easy for the respondents as a whole, consistent with findings by (Hidayat et al., 2025)

This indicates that although the instrument demonstrates adequate validity and reliability, its items are concentrated within a single difficulty level and fail to represent the full spectrum of student ability. Therefore, future development of this instrument should focus on broadening the range of item difficulty, particularly by adding more challenging and easier items, so that it can more comprehensively capture variations in students' mathematical reasoning ability. The Wright Map provides an important implication beyond the descriptive comparison of person ability and item difficulty. The slightly higher person

mean indicates that, on average, students possessed reasoning ability that was somewhat greater than the level of challenge provided by the instrument. This mismatch suggests that the current item set has limited sensitivity for distinguishing students at the upper end of the ability continuum. In practical terms, high-ability students may obtain relatively high scores without being sufficiently challenged, resulting in reduced measurement precision around the upper ability range. At the same time, the absence of items below −1 logit limits the instrument's capacity to differentiate students with lower reasoning ability. Thus, the Wright Map indicates that future instrument development should not merely increase the number of items, but should deliberately construct items across a wider difficulty continuum. Additional high-difficulty items should require more complex justification, multi-step reasoning, or evaluation of alternative mathematical arguments, whereas lower-difficulty items could focus on fundamental reasoning processes. Such a balanced distribution would improve the instrument's targeting and provide more precise estimates across different levels of students' mathematical reasoning ability

2. Discussion

Findings from this research indicate the quality of the developed mathematical reasoning instrument on linear equations. First, all four items were confirmed valid through Rasch criteria, with Outfit MNSQ values (0.71–0.97) indicating productive response patterns free from erratic behavior(Peeters & Augustine, 2023). The strong Point Measure Correlations (0.82–0.84) confirm each item consistently discriminates between students of higher and lower reasoning ability, echoed by (Rusyid et al., 2024) in their algebraic thinking instrument validation.

Second, the Person Reliability of 0.87 with a separation index of 2.63 indicates sufficient power to distinguish at least two meaningfully different ability groups. This is particularly valuable for formative assessment, as teachers can use such stratification to tailor instruction. The result surpasses the 0.80 benchmark noted by (Hidayat et al., 2025), strengthening confidence in the instrument's reliability. The importance of reliable reasoning instruments for junior high school students has been consistently emphasized in recent studies, including (Siregar et al., 2025), who both highlight the need for diagnostically precise tools.

Third, the difficulty analysis reveals a critical limitation: all four items fall within the moderate difficulty band (0.00 to +0.70 logits), with no items classified as difficult or easy. While appropriate for measuring average-level students, this creates a ceiling effect for high-ability learners and a floor effect for lower-ability learners. (Qirom & Nurlaelah, 2023) similarly observed that instruments with narrow item difficulty ranges fail to capture the full performance spectrum. The Wright Map confirms this gap: respondents distributed across a wide range were not adequately matched by items at the extreme ends.

Fourth, the narrow item pool (four items) inherently limits diagnostic precision. (Darmana et al., 2023) support the argument that Rasch-based instruments benefit from larger item pools to improve coverage. Future development should add items targeting difficult and easy categories while maintaining polytomous scoring to preserve diagnostic richness. Differential Item Functioning (DIF) analysis was not included in the present study because the research was designed as an initial instrument calibration involving a relatively small sample of 30 students from a single school. The available sample and single-school setting limited the feasibility and interpretability of subgroup comparisons. Nevertheless, DIF analysis is important in subsequent studies because it can identify whether students from different groups, such as male and female students or students from different school contexts, with comparable levels of mathematical reasoning ability have different probabilities of responding successfully to the same item. Conducting DIF analysis would therefore provide additional evidence of item fairness and ensure that item difficulty reflects the intended mathematical reasoning construct rather than characteristics associated with particular student groups or contexts.

From a pedagogical perspective, the moderate difficulty of all items suits formative assessment of average-performing students. The use of NCTM's four reasoning indicators ensures content validity and alignment with internationally recognized standards. The persistent difficulties students face in algebraic reasoning, particularly in tasks requiring logical inference and justification (Dwi Ikasari & Firmansyah, 2023) and (Asoraya & Ruli, 2023) reinforce the importance of having validated instruments that can accurately diagnose reasoning deficits. (Hadi et al., 2024) further argue that such instruments are essential for developing contextually appropriate mathematics instruction for junior high school students.

In summary, this study contributes to the growing body of literature on Rasch-based mathematics instrument development in Indonesia and demonstrates the effectiveness of Ministep software for polytomous item calibration. The instrument is suitable for formative assessment of mathematical reasoning on linear equations at the junior high school level, with future revisions addressing identified gaps in item difficulty coverage.

CONCLUSION

This study analyzed the quality of a four-item mathematical reasoning ability instrument on linear equations for junior high school students using Rasch Model analysis via Ministep. All four items were declared valid, meeting Outfit MNSQ (0.71–0.97), Outfit ZSTD (−0.34 to 0.15), and Point Measure Correlation (0.82–0.84) criteria. Instrument reliability was 0.87 with a separation index of 2.63, classified as good. All items were in the moderate difficulty range (logit: 0.00 to +0.70), with no items classified as difficult or easy. Wright Map analysis revealed that average student ability exceeded average item difficulty, indicating a ceiling effect for high-ability students and the need to broaden item difficulty coverage in future development.

The instrument is suitable for competency-based formative assessment of mathematical reasoning on linear equations. Future research should expand the item pool to include difficult and easy items, involve multi-school samples, and apply Differential Item Functioning (DIF) analysis to examine item fairness across relevant student groups and school contexts.

ACKNOWLEDGMENTS

The authors thank the principal, mathematics teachers, and Grade 8 students of the participating school in Cirebon for their cooperation during data collection. No external funding was received for this research.

REFERENCES

Ambarwati, D., & Maarif, S. (2025). Rasch Model Analysis: Development of Students' Numeracy Literacy Interest Instruments in Mathematics Learning. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, *11*(1), 41–52. https://doi.org/10.29100/jp2m.v11i1.7279

Aprilia, N., Setiani, Y., & FS, C. A. H. (2023). Pengembangan Instrumen Tes Numerasi Pada Asesmen Kompetensi Minimum Yang Bernilai Budaya Lokal. *Jurnal Educatio FKIP UNMA*, *9*(2), 850–857. https://doi.org/10.31949/educatio.v9i2.4824

Asoraya, M. S., & Ruli, R. M. (2023). Analisis Kemampuan Penalaran Matematis Siswa SMP pada Materi Relasi dan Fungsi. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *7*(3), 3053–3066. https://doi.org/10.31004/cendekia.v7i3.2412

Darmana, A., Pulungan, A. N., Nasution, H. A., Nst, M. A., & Faradilla, P. (2023). Test Instrument Validation with Rasch Model. *Proceedings of the 5th International Conference on Innovation in Education, Science, and Culture*. https://doi.org/10.4108/eai.24-10-2023.2342115

Dwi Ikasari, & Firmansyah, F. (2023). Pengaruh Model Pembelajaran Problem Based Learning Berbantuan Lembar Kerja Peserta Didik (LKPD) terhadap Kemampuan Penalaran Matematis Siswa di MTs. Lab Ikip Al-Washliyah Medan. *Jurnal Ilmiah Pendidikan Matematika Al Qalasadi*, *7*(1), 55–65. https://doi.org/10.32505/qalasadi.v7i1.5882

Hadi, S., Rahman, A., & Gafur, A. (2024). Students' Mathematical Reasoning Ability in Junior High School Mathematics Learning with Contextual Reasoning. *Journal of Education and Social Science*, *1*(1), 6–11. https://doi.org/10.70716/jees.v1i1.57

Hidayat, R., Dahlan, J. A., & Nurlaelah, E. (2025). *Rasch Model Analysis of Algebraic Numeracy Literacy Items Using Winsteps*. *4*(4), 2309–2321. https://doi.org/10.58421/gehu.v4i4.736

Istiqomah, S., Lusiyana, D., & Hidayat, R. (2025). *Rasch Model Analysis of Spatial Visualization Test Items on the Pythagorean Theorem for Junior High School Students Using Ministep*. *4*(3), 931–941.

Mukuka, A., Balimuttajjo, S., & Mutarutinya, V. (2023). Teacher efforts towards the development of students' mathematical reasoning skills. *Heliyon*, *9*(4). https://doi.org/10.1016/j.heliyon.2023.e14789

NCTM, N. C. (2000). Principles and standards for school mathematics. *Reston, VA: NCT M*.

Peeters, M. J., & Augustine, J. M. (2023). Using Rasch measurement for instrument rating scale refinement. *Currents in Pharmacy Teaching and Learning*, *15*(1), 110–118. https://doi.org/10.1016/j.cptl.2023.02.015

Qirom, M. S., & Nurlaelah, E. (2023). Rasch Model: Analyzing the Items Quality of Mathematics HigherOrder Thinking Skill Instrument. *JPP (Jurnal Pendidikan Dan Pembelajaran)*, *30*(1), 8–16. https://doi.org/10.17977/um047v30i12023p8-16

Rusyid, H. K., Suryadi, D., Herman, T., Adnan, M., Lutfi, A., & Mukhibin, A. (2024). Rasch modelling approach to measure the quality of algebraic thinking test item for junior high school students. *Beta: Jurnal Tadris Matematika*, *17*(1), 44–58. https://doi.org/10.20414/betajtm.v17i1.652

Sajdah, S. N., & Hidayati, K. (2026). Development of a Diagnostic Assessment Instrument to Measure the Mathematical Reasoning Ability of Junior High School Students. *Al Qalam: Jurnal Ilmiah Keagamaan Dan Kemasyarakatan*, *20*(3), 1613–1629. https://doi.org/10.35931/aq.v20i3.5810

Sarifah, I., Kuswati, E., Suniasih, W., Fahrurrozi, F., & Susilo, G. (2024). Cognitive test instruments to measure student mathematics ability in elementary school: Rasch analysis. *Educational Administration: Theory and Practice*, *30*(4), 767–779. https://doi.org/10.53555/kuey.v30i4.1557

Siregar, E. Y., Elindra, R., & Sikumbang, V. M. S. (2025). Pengembangan instrumen penilaian untuk mengukur kemampuan penalaran matematis siswa kelas VIII SMP N 2 Padangsidimpuan. *JURNAL MathEdu (Mathematic Education Journal)*, *8*(2), 46–55.

Sumintono, B. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Wardani, R. T., Nuraeni, E., & Diana, S. (2025). Rasch model analysis of essay questions to measure literacy and numeracy skills in plant and animal bioprocess topics based on AKM. *REID (Research and Evaluation in Education)*, *11*(1), 29–44. https://doi.org/10.21831/reid.v11i1.85614

Wiyarsi, A., Prodjosantoso, A. K., & Nugraheni, A. R. E. (2020). Students' chemical literacy level: A case on electrochemistry topic. *Journal of Physics: Conference Series*, *1440*(1), 12019. https://doi.org/10.1088/1742-6596/1440/1/012019