

Development and Rasch Validation of a Mathematical Communication Skills Test for Indonesian Elementary School Students

Ahmad Rustam

ahmadrustam@un-sultra.ac.id

Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sulawesi Tenggara

Zibar L. Chairan Parisu

Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sulawesi Tenggara

Hesti Hesti

Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sulawesi Tenggara

Nisraeni Nisraeni

Universitas Cokroaminoto Palopo

Deni Iriyadi

Program Pascasarjana, Universitas Islam Negeri Sultan Maulana Hasanuddin Banten

Research Article

Keywords: Rasch measurement model, mathematical communication skills, primary mathematics assessment, item response theory, educational measurement practice

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10539270/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Valid assessment of mathematical communication is needed to diagnose elementary students' ability to represent, model, and explain mathematical ideas. This study developed and validated a cognitive test for Indonesian elementary students using the Rasch measurement model. A research and development design was conducted through preliminary analysis, blueprint and item construction, expert review, field testing, and Rasch calibration. The trial involved 529 fifth- and sixth-grade students from three schools in Kendari City and South Konawe Regency, Southeast Sulawesi, Indonesia. The 25-item multiple-choice test measured three indicators: representing situations through pictures or tables, translating contextual situations into mathematical models or symbols, and selecting appropriate written explanations. Responses were analyzed with WINSTEPS 3.73 to examine dimensionality, item fit, reliability, separation, item difficulty, targeting, differential item functioning by gender and grade, and test information. Results showed a dominant dimension, with 32.9% raw variance explained and a first residual contrast of 5.3%. Internal consistency was excellent ($KR-20 = .92$), person reliability was satisfactory (.81), and item reliability was high (.94). Most items fitted the model, although two difficult items need review. No gender or grade DIF was detected. The instrument provides preliminary validity, reliability, and fairness evidence for elementary mathematics assessment in Indonesian primary classroom contexts.

1. Introduction

Educational measurement is central to instructional decision-making because the quality of an assessment instrument determines the credibility of the inferences made from student scores. Contemporary validity theory does not treat validity as a property of a test itself; rather, validity refers to the degree to which empirical evidence and theoretical rationale support the intended interpretation and use of scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME],

(Kane, 2013; Messick, 1989). In classroom contexts, valid score interpretation is required for diagnosing learning needs, monitoring curriculum outcomes, and designing targeted instructional support.

For many years, Classical Test Theory (CTT) was the dominant framework for test development. CTT remains useful for estimating total-score reliability and sample-level item statistics, but its item difficulty and discrimination indices are dependent on the specific examinee sample, while examinee scores are dependent on the particular set of administered items. Item Response Theory (IRT), particularly the Rasch measurement model, provides a stronger framework for instrument validation because person ability and item difficulty are placed on a common interval logit scale

(Bond and Fox, 2015; Boone et al., 2000; Embretson and Reise, 2013). Rasch analysis also enables researchers to evaluate dimensionality, item fit, item difficulty hierarchy, reliability, separation, differential item functioning (DIF), and the alignment between item difficulty and student ability through Wright maps.

Mathematical communication is a core competency in mathematics learning. It includes students' capacity to represent mathematical situations visually or tabularly, translate contextual problems into mathematical symbols or models, and communicate solution processes through written explanations. International mathematics education documents emphasize communication and representation as essential mathematical processes that support reasoning, problem solving, and conceptual understanding (American Educational Research Association et al., 2014; National Council of Teachers of Mathematics, 2000). These abilities are especially important at the elementary level because they form the basis for later mathematical literacy and for students' ability to connect real-world contexts with formal mathematical expressions.

Despite its importance, students frequently struggle to communicate mathematical ideas, especially when they must coordinate diagrams, symbolic expressions, real-world contexts, and written reasoning. Within the Indonesian elementary curriculum context, valid and fair instruments that specifically measure mathematical communication remain limited, particularly instruments that have been empirically calibrated using Rasch measurement. Many existing classroom tests emphasize procedural correctness but provide limited evidence about construct representation, item-person targeting, measurement precision, and item fairness across student groups.

Previous Rasch-based educational studies demonstrate that the model is useful for examining item quality, score reliability, construct structure, and fairness in cognitive and affective instruments. However, fewer studies have integrated content alignment, unidimensionality analysis, item fit, difficulty distribution, reliability, separation, DIF by gender and grade, Wright map interpretation, information function analysis, and indicator-level achievement in a single validation framework for elementary mathematical communication. Addressing this gap is important because teachers require assessments that are not only aligned with curricular indicators but also psychometrically defensible and useful for instructional decisions.

Accordingly, this study aimed to develop and validate a cognitive test instrument for measuring elementary school students' mathematical communication skills using the Rasch measurement model. The study was guided by four research objectives: (1) to construct items aligned with curriculum-based indicators of mathematical communication; (2) to evaluate psychometric quality through dimensionality, item fit, reliability, separation, difficulty, targeting, and information function analyses; (3) to examine item fairness through gender- and grade-based DIF analysis; and (4) to describe student achievement across mathematical communication indicators. The contribution of this study lies in its integrated Rasch-based validation approach and its focus on elementary mathematical communication within an Indonesian curriculum context.

2. Theoretical and Measurement Framework

2.1 Mathematical communication as an assessment construct

Mathematical communication is broadly recognized as students' capacity to express, translate, and explain mathematical ideas through visual, symbolic, and written representations. This conceptualization aligns with established frameworks asserting that mathematics learning encompasses not only computation but also representation, reasoning, communication, and connection across multiple forms of mathematical expression (Bossé et al., 2010; Cormas and Middlemiss, 2019; Irving and Bell, 2004). The National Council of Teachers of Mathematics (NCTM) has consistently emphasized communication as a core process standard through which students organize and consolidate mathematical thinking (Barham, 2020; Bybee et al., 1997; Eddy et al., 2015). Similarly, the National Research Council has underscored that mathematical proficiency requires interconnected strands including conceptual understanding, strategic competence, and adaptive reasoning (Baker, 2000; Eddy et al., 2015; Sorto, 2011).

The operationalization of mathematical communication through observable indicators—representing contextual situations visually, translating situations into mathematical models, and explaining ideas in written language—reflects the multidimensional nature of the construct. Research confirms that students must be able to represent mathematical situations in different ways and understand how different representations serve different purposes (Colonnese, 2020; Koirala and Bowman, 2003). Mathematical communication thus functions as both a cognitive and social process enabling knowledge construction (Bybee et al., 1997; Genç et al., 2025).

Although multiple-choice items cannot fully capture open-ended communication, they can provide efficient evidence of students' ability to recognize appropriate representations, models, and explanations when carefully constructed and aligned with the intended construct (Asmida et al., 2018; Mohamed et al., 2023). Such formats are particularly useful for assessing whether students can identify correct visual representations, symbolic models, and written reasoning within structured response options, provided items are rigorously developed to target specific communication indicators.

2.2 Rasch measurement and validity evidence

The Rasch model specifies the probability of a correct response as a logistic function of the difference between person ability and item difficulty, expressing both parameters on a common logit scale that enables direct interpretation of item-person targeting and the hierarchy of item difficulty (Malonisio and Malonisio, 2023; McNamara and Knoch, 2012; Pérez-Chada et al., 2024). This co-calibration of persons and items on a single continuum constitutes the model's central measurement advantage, distinguishing it from classical test theory approaches that rely on sample-dependent statistics (Nick et al., 2021; Prior et al., 2018; Sorrentino et al., 2021). The Rasch framework has been widely applied across educational and psychological measurement contexts, including language testing (X. Fan, 1998; Hammond et al., 2023), achievement testing (Hammond et al., 2023; Wong et al., 2021), and health outcome measurement (Klerk, 2022; Manz et al., 2014), demonstrating its versatility in producing interval-level measures from ordinal data.

Within a validation framework, Rasch-derived evidence should not be interpreted in isolation but must be integrated with multiple sources of validity evidence—including content evidence, response process evidence, internal structure, reliability, fairness, and consequences for score use (J. Fan and Knoch, 2019; Raman et al., 2024). The *Standards for Educational and Psychological Testing* outline these complementary sources as essential components of a coherent validity argument. Kane's argument-based approach to validation further emphasizes that the strength of validity claims depends on the integration of diverse evidence strands (J. Fan and Knoch, 2019; Suryadi et al., 2023).

For the present instrument, validity evidence was gathered through expert review, dimensionality analysis, item fit statistics, reliability and separation indices, item difficulty distribution, Wright map targeting, test information function, DIF analysis, and indicator-level achievement. Such comprehensive approaches—combining Rasch analysis with content validation, dimensionality testing, and bias detection—have been employed across diverse measurement contexts (Deeba and Ahmad, 2023; Eslahi et al., 2021; Malonisio and Malonisio, 2023; Nick et al., 2021). This multi-faceted framework supports a coherent argument that the resulting scores reflect mathematical communication ability rather than construct-irrelevant variance.

Table 1
Instrument blueprint and construct mapping

Construct domain	Operational indicator	Assessment evidence	Item numbers
Visual or tabular representation	Representing a contextual situation in a picture or table	Students select or interpret a representation that matches a mathematical situation.	1, 2, 3, 5, 6, 7, 8, 9, 10
Symbolic modeling	Translating a situation into a mathematical model or symbol	Students identify a suitable mathematical expression, equation, or symbolic relation.	4, 11, 12, 13, 14, 15, 17, 18, 19, 21, 23, 25
Written mathematical explanation	Explaining mathematical ideas or solution processes in written language	Students select statements that correctly explain reasoning, solution steps, or the meaning of a model.	16, 20, 22, 24
Integrated communication	Combining visual, symbolic, and explanatory elements	Items require students to coordinate two or more forms of mathematical communication.	Several contextual items across the instrument

3. Methods

This study employed a research and development (R&D) design focused on producing and validating a cognitive assessment instrument (Amira and Wulan, 2021). The development process consisted of five stages: (1) preliminary study, (2) construction of the test blueprint and items, (3) expert validation, (4) field testing, and (5) empirical calibration using the Rasch measurement model (Putri et al., 2020;; (Ridwan et al., 2023). The validation logic was organized to generate evidence related to content

relevance, construct alignment, internal structure, item functioning, reliability, fairness, and score interpretation (Hrnjičić and Alihodžić, 2024; Planinić et al., 2019).

The field trial involved 529 elementary school students from three partner schools: SDN 34 Kendari, SDN 4 Ranomeeto, and SDN 88 Kendari, representing Kendari City and South Konawe Regency, Southeast Sulawesi, Indonesia. The participants consisted of 265 male and 264 female students; 258 were in Grade 5 and 271 in Grade 6. This distribution provided adequate response data for Rasch calibration and for examining item functioning across gender and grade groups (Ridwan et al., 2023; Motta et al., 2020; .

The instrument measured mathematical communication skills according to elementary school curriculum standards (Bellini et al., 2019; Fritz-Stratmann et al., 2014). The final field-test version consisted of 25 multiple-choice items, each mapped to at least one indicator of mathematical communication (Amira and Wulan, 2021; Veith et al., 2022). Initial items were developed based on the test blueprint, basic competencies, and intended cognitive processes (Chin et al., 2022). Expert validation was conducted by mathematics education lecturers and elementary school teachers, focusing on content relevance, construct alignment, clarity, linguistic appropriateness, distractor quality, and indicator consistency (Hrnjičić and Alihodžić, 2024; Mešić et al., 2019). Items were revised according to expert feedback before field testing, providing content validity evidence prior to empirical calibration (Novak et al., 2022; J. W. Pellegrino et al., 2001).

The field test was administered after ethical approval had been obtained from the Research Ethics Committee of the Institute for Research and Community Service, Universitas Sulawesi Tenggara, Indonesia (Approval No. 66/R/09/LPPM/VIII/2025; July 1, 2025), permission had been obtained from the participating schools, and written informed consent had been obtained from parents or legal guardians. Students were informed about the assessment purpose, response confidentiality, voluntary participation, and their right to withdraw. All procedures minimized burden and protected student welfare, in accordance with the approved protocol and relevant ethical principles for research involving human participants.

Student responses were analyzed using WINSTEPS version 3.73 (Qudratuddarsi et al., 2022; Syawaludin et al., 2019). Unidimensionality was evaluated using principal component analysis of Rasch standardized residuals, with raw variance explained by measures of at least 20% and residual contrasts below 15% treated as evidence supporting a dominant measurement dimension (Putri, 2023; Qudratuddarsi et al., 2022; Syawaludin et al., 2019). Item fit was evaluated using infit and outfit mean square (MNSQ), standardized fit (ZSTD), and point-measure correlation (Dwiliesanti and Yudiarso, 2022; Idulfilastri et al., 2021; Planinić et al., 2019). MNSQ values between 0.50 and 1.50 were interpreted as productive for measurement (Dwiliesanti and Yudiarso, 2022; Segar, 2024), while ZSTD values were used cautiously due to sensitivity to large sample sizes (Planinić et al., 2019; Tunç, 2023). Point-measure correlations were expected to be positive, indicating items moved in the intended construct direction (Qudratuddarsi et al., 2022). Reliability evidence was obtained from KR-20/Cronbach's alpha, person reliability, item reliability, and separation indices (Chin et al., 2022; J. L. Pellegrino et al., 2020; J. W. Pellegrino et al.,

2001; Wijaya, 2023). DIF analysis was conducted by gender and grade, with items at $p < .05$ treated as potentially flagged, though practical significance was interpreted cautiously (Motta et al., 2020; Ridwan et al., 2023; Антипкина and Ludlow, 2020). Wright maps and the test information function evaluated item-person targeting and measurement precision across the ability continuum (Arias et al., 2015; Planinić et al., 2019; Tunç, 2023).

4. Results

4.1 Participant profile and instrument structure

The sample consisted of 529 students, with an almost equal gender distribution: 265 male students and 264 female students. The grade-level distribution was also balanced, with 258 Grade 5 students and 271 Grade 6 students. The final trial instrument consisted of 25 items mapped to mathematical communication indicators. The sample size and group distribution were sufficient for the intended Rasch calibration and exploratory DIF analysis.

4.2 Unidimensionality evidence

Unidimensionality analysis was conducted to examine whether the instrument primarily measured one dominant construct, namely mathematical communication skills. The results showed that the raw variance explained by Rasch measures was 32.9% empirically and 33.5% in the modeled estimate. The unexplained variance in the first to fifth residual contrasts ranged from 3.3% to 5.3%, all below the 15% threshold. These findings support the interpretation that the instrument measures a dominant construct and is appropriate for subsequent Rasch analysis. However, because the first residual contrast had an eigenvalue of 2.0, future studies should continue examining whether specific item clusters represent meaningful subdomains of mathematical communication.

Table 2
Principal component analysis of Rasch standardized residuals

Component	Eigenvalue	Empirical %	Modeled %
Total raw variance in observations	37.3	100.0	100.0
Raw variance explained by measures	12.3	32.9	33.5
Raw variance explained by persons	7.5	20.0	20.4
Raw variance explained by items	4.8	12.9	13.2
Raw unexplained variance (total)	25.0	67.1	66.5
Unexplained variance in 1st contrast	2.0	5.3	7.9
Unexplained variance in 2nd contrast	1.7	4.7	7.0
Unexplained variance in 3rd contrast	1.5	4.1	6.1
Unexplained variance in 4th contrast	1.4	3.6	5.4
Unexplained variance in 5th contrast	1.2	3.3	4.9

4.3 Item difficulty and test targeting

Rasch item difficulty analysis placed the 25 items on a logit scale. Item 17 was the most difficult item (2.01 logits), indicating that it was most likely to be answered correctly only by students with relatively high mathematical communication ability. The remaining 24 items were located between -0.83 and $+0.83$ logits, which falls within the moderate difficulty range. No item was categorized as easy or very easy. This distribution indicates that the instrument is well suited to measuring middle-to-high ability students but may provide less precise information for students with lower ability.

Table 3
Distribution of item difficulty levels

Difficulty category	Logit criterion	Items	Number of items	Percentage
Very difficult	$> +1.00$	17	1	4%
Moderate	-1.00 to $+1.00$	1–16, 18–25	24	96%
Easy	< -1.00 to -2.00	-	0	0%
Very easy	< -2.00	-	0	0%

4.4 Item fit

Item fit analysis was used to evaluate whether each item functioned consistently with the Rasch model. Overall, the instrument demonstrated acceptable item functioning because all items had positive point-measure correlations and most items showed MNSQ values within the productive range. However, two high-difficulty items, items 17 and 24, showed elevated outfit statistics. These items were retained in the

trial analysis because they were theoretically aligned with the construct and helped represent more challenging mathematical communication tasks, but they should be substantively reviewed before operational use. This interpretation is more appropriate than relying on ZSTD alone because ZSTD values can be inflated in large samples.

Table 4
Summary of item fit evidence

Fit evidence	Result	Interpretation
Point-measure correlation	All items positive; range approximately .42-.63	All items moved in the expected direction of the construct.
Infit MNSQ	Most items within .50 – 1.50; item 17 = 1.55 and item 24 = 1.52	Generally productive; items 17 and 24 require review.
Outfit MNSQ	Most items within .50 – 1.50; item 17 = 2.30 and item 24 = 1.81	Unexpected responses were detected on two high-difficulty items.
ZSTD	Several values outside – 2 to + 2	Used only as supporting evidence because of sample-size sensitivity.
Decision	25 items retained for trial interpretation	Retain with revision notes for difficult items.

4.5 Reliability, Separation, and Global Model Fit

Reliability analysis indicated that the instrument had strong internal consistency. The KR-20/Cronbach's alpha value was .92, suggesting excellent raw-score reliability. Person reliability was .81 for non-extreme respondents and .75 when extreme and non-extreme respondents were included. Item reliability was .94, indicating that the item difficulty hierarchy was stable and well estimated. The item separation index of 3.94–4.05 showed that the instrument could distinguish several levels of item difficulty. The person separation index was lower but still useful for identifying broad differences in student ability. Average item fit statistics were close to the expected Rasch value of 1.00, supporting global model-data agreement.

Table 5
Reliability and separation summary

Statistic	Value	Interpretation
KR-20/Cronbach's alpha	.92	Excellent internal consistency.
Person reliability (non-extreme)	.81	Good consistency of person ordering.
Person reliability (all respondents)	.75	Acceptable when extreme cases are included.
Person separation (non-extreme)	2.07	Able to distinguish broad ability groups.
Item reliability	.94	High stability of item hierarchy.
Item separation	3.94–4.05	Strong spread of item difficulty estimates.
Mean person measure	+ 1.26 logits	Average student ability exceeded average item difficulty.
Mean item measure	0.00 logits	Item scale anchored at the Rasch mean.
Mean item infit/outfit MNSQ	1.00 / .95	Close to expected model values.

4.6 Differential item functioning

DIF analysis was conducted to evaluate whether items functioned differently across gender and grade groups after controlling for overall ability. The analysis did not flag any of the 25 items as biased by gender or grade because the probability values did not meet the significance criterion of $p < .05$. This result provides preliminary fairness evidence for the instrument across the observed student groups. Nevertheless, future studies should also report DIF contrast magnitudes and apply multiplicity-adjusted interpretations so that statistical significance and practical relevance can be considered together.

4.7 Wright map and test information function

The Wright map placed student ability and item difficulty on the same logit scale. The map showed that high-ability students, including those between approximately + 3.00 and + 4.00 logits, were able to answer the most difficult item (item 17). Students around + 2.00 logits were generally able to answer items with moderate difficulty. The average person measure (+ 1.26 logits) was higher than the item mean, indicating that the instrument was somewhat easier than the average ability of the sample, although it still contained challenging items for high-ability students. The map also revealed that easier items should be added to improve targeting for students at the lower end of the ability distribution.

The test information function indicated that the instrument provided the most accurate information in the – 3 to + 3 logit range. This is desirable for a heterogeneous elementary student population because it supports interpretation across low, moderate, and high ability levels. The curve was centered around the middle range of ability, showing that the instrument was particularly informative for students near the average ability level while still providing useful information across a broad logit span.

4.8 Indicator-level achievement

The item indicators were further analyzed to describe the percentage of correct answers across mathematical communication domains. The highest achievement was found in items requiring students to use mathematical symbols and pictures in story contexts (78.79%). The lowest achievement was observed in items requiring students to state mathematical models and explain solution processes (67.11%). These findings indicate that students were more successful with simple symbolic and visual representations than with tasks requiring more complex modeling and written explanation.

Table 6
Achievement by mathematical communication indicator

Mathematical communication indicator	Item numbers	Correct answers (%)
Using mathematical symbols and pictures in the context of a story	1, 2, 5, 6, 7	78.79
Writing a mathematical model and explaining the solution process	4, 12, 14	77.50
Drawing situations and stating mathematical relationships	3, 8, 9, 10	75.71
Writing symbolic expressions and giving reasons or explanations	11, 13, 15, 18, 21	73.04
Stating the mathematical model and explaining the process	17, 19, 23, 25	67.11
Drawing, writing mathematical models, and explaining meaning	16, 20, 22, 24	71.41

5. Discussion

5.1 Validity evidence and psychometric quality

The results provide multiple sources of validity evidence for the developed instrument. The dimensionality analysis supports the interpretation that the items measure a dominant construct of mathematical communication skills. The reliability results also indicate that the instrument is sufficiently stable for educational assessment purposes. These findings are consistent with Rasch measurement principles, which emphasize the joint evaluation of item difficulty, person ability, item fit, reliability, separation, and construct structure (Bond and Fox, 2015; Boone et al., 2000).

The use of Rasch analysis strengthened the validation argument because the analysis did more than estimate internal consistency. It examined whether items functioned in the expected direction, whether the item hierarchy was stable, whether the test targeted the observed student ability distribution, and whether items were fair across gender and grade groups. This type of integrated evidence is important for defensible score interpretation in educational assessment (Kane, 2013; Messick, 1989).

5.2 Interpretation of item difficulty and targeting

The item difficulty distribution suggests that the instrument is most appropriate for students with moderate to high mathematical communication ability. The presence of one very difficult item is useful because it helps distinguish high-ability students, but the absence of easy and very easy items limits precision for students at the lower end of the ability scale. This finding is consistent with the Wright map, where the mean person measure was higher than the mean item difficulty.

For operational use, additional lower-difficulty items should be developed. These items should target foundational communication tasks, such as matching a simple picture to a story situation, identifying a single-step mathematical expression, reading a basic table, or selecting a short explanation of a simple computation. Adding easier items would broaden the measurement range and improve the instrument's usefulness for diagnosing students with emerging mathematical communication skills.

5.3 Fairness and bias evidence

The absence of statistically significant DIF by gender and grade provides preliminary evidence that the instrument functioned similarly across the observed student groups. In practical terms, students with comparable mathematical communication ability had similar chances of answering the items correctly regardless of gender or grade level. This finding strengthens the fairness argument for the instrument. However, DIF results should not be overinterpreted because the analysis reported only probability values. Future studies should report DIF contrast magnitudes, confidence intervals, and adjusted significance criteria to distinguish statistical from practical bias.

5.4 Instructional implications

The indicator-level findings have direct implications for mathematics instruction. Students performed relatively well on items involving symbols and pictures in story contexts but performed less well on items requiring mathematical modeling and explanation of solution processes. This pattern suggests that teachers should provide more structured opportunities for students to write mathematical explanations, justify reasoning, connect diagrams with symbolic expressions, and interpret the meaning of calculations. Instruction should not stop at obtaining correct answers; it should also require students to communicate why a representation or model is appropriate.

The instrument can be used as a diagnostic tool to identify which aspects of mathematical communication require instructional support. For example, lower achievement in modeling and written explanation indicates a need for learning activities that ask students to translate everyday situations into equations and explain solution steps in their own words. Such activities are consistent with international recommendations that mathematical communication and representation should be integrated into mathematics learning rather than treated as supplementary skills (American Educational Research Association et al., 2014; Bossé et al., 2010; National Research Council, 2001).

5.5 Limitations and directions for future research

First, the sample was drawn from three elementary schools in two regions of Southeast Sulawesi; therefore, broader validation across provinces, school types, and socio-economic contexts is needed before national-scale use.

Second, the instrument used multiple-choice items. Although this format can measure selected aspects of mathematical communication, it does not fully capture open-ended written reasoning. Future studies should combine multiple-choice items with constructed-response tasks and analytic scoring rubrics.

Third, items 17 and 24 showed elevated outfit statistics. These items should be reviewed through cognitive interviews, distractor analysis, and content analysis before operational deployment.

Fourth, future studies should examine local independence, criterion-related validity, test-retest reliability, and score interpretation using external criteria such as teacher ratings, classroom performance, or standardized mathematics achievement data.

Fifth, DIF contrast magnitudes were not reported in the present analysis. Future research should include both statistical and practical DIF criteria.

6. Conclusion

This study developed and validated a 25-item cognitive test instrument for measuring elementary school students' mathematical communication skills using the Rasch measurement model. The items were aligned with curriculum standards and indicators of mathematical communication, including visual or tabular representation, mathematical modeling or symbolic expression, and written explanation. Rasch analysis provided evidence of a dominant measurement dimension, high internal consistency, satisfactory person reliability, high item reliability, adequate item-person targeting, and no statistically detected gender or grade DIF. The Wright map and information function showed that the instrument was most informative for students within the -3 to $+3$ logit range, although additional easier items are needed to improve precision for lower-ability students. Indicator-level findings showed that students were relatively stronger in simple symbolic and visual representation but weaker in modeling and written explanation. Overall, the instrument offers preliminary evidence of validity, reliability, and fairness for assessing mathematical communication skills in elementary education and can serve as a reference for further Rasch-based mathematics assessment development.

Declarations

Acknowledgements

The authors thank the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for supporting this research. The authors also acknowledge Universitas Sulawesi Tenggara, the participating schools, teachers, and students for their cooperation.

Funding

This research was funded by a research grant awarded to Universitas Sulawesi Tenggara under Grant No. 070/C3/AL.04/2025.

Author contributions

Author contributions were defined according to the CRediT taxonomy. Ahmad Rustam contributed to conceptualization, methodology, supervision, validation, investigation, writing, visualization, project administration, and funding acquisition. Chairan Zibar L. Parisu and Hesti contributed to methodology, validation, investigation, data curation, and writing. Nisraeni contributed to software, formal analysis, data curation, visualization, and writing. Deni Iriyadi contributed to methodology, formal analysis, validation, review, and editing. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This study involved human participants (elementary school students). The study protocol was reviewed and approved by the Research Ethics Committee of the Institute for Research and Community Service, Universitas Sulawesi Tenggara, Indonesia (Approval No. 66/R/09/LPPM/VIII/2025; date of approval: July 1, 2025). All procedures involving human participants were performed in accordance with the institutional ethical guidelines of Universitas Sulawesi Tenggara, relevant national guidelines and regulations, and the ethical principles of the Declaration of Helsinki and its later amendments.

Consent to participate

Written informed consent to participate was obtained from the parents or legal guardians of all participating students before data collection. Students were informed about the purpose of the assessment, the voluntary nature of participation, confidentiality and anonymity, and their right to refuse or withdraw from participation without penalty.

Consent to publish

Not applicable.

Data availability

The data are available from the corresponding author upon reasonable request, subject to ethical, institutional, and confidentiality restrictions. Eligible research materials may also be made available after the publication embargo period at: <https://ahmadrustam.ardenjaya.com/penelitian/>

AI-assisted language editing statement

An AI-assisted writing tool was used only for language editing, academic style refinement, grammatical correction, and manuscript organization. The authors reviewed and approved all content and remain fully responsible for its accuracy, originality, and integrity.

References

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association; 2014.
2. Amira N, Wulan AR. (2021). *Development of Written Test Instruments to Measure Basic Skills of STEAM (Science, Technology, Engineering, Arts, and Mathematics) Junior High School Students on Science Concept*. <https://doi.org/10.2991/assehr.k.210203.088>
3. Asmida A, Sugiatno S, Hartoyo A. Developing the Mathematics Conceptual Understanding and Procedural Fluency Through Didactical Anticipatory Approach Equipped With Teaching AIDS. *Jetl (Journal Educ Teach Learning)*. 2018. <https://doi.org/10.26737/jetl.v3i2.796>.
4. Baker JL. Evaluating the Impact of Development Projects on Poverty A Handbook for Practitioners. World Bank; 2000.
5. Barham AI. Exploring in-Service Mathematics Teachers' Perceived Professional Development Needs Related to the Strands of Mathematical Proficiency (SMP). *Eurasia J Math Sci Technol Educ*. 2020. <https://doi.org/10.29333/ejmste/8399>.
6. Bellini D, Crescentini A, Zanolla G, Cubico S, Favretto G, Faccincani L, Ardolino P, Giancesini G. (2019). Mathematical Competence Scale (MCS) for Primary School: The Psychometric Properties and the Validation of an Instrument to Enhance the Sustainability of Talents Development Through the Numeracy Skills Assessment. *Sustainability*. <https://doi.org/10.3390/su11092569>
7. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Routledge; 2015.
8. Boone WR, Jones JM, Shapiro SS. Using videotaped specimens to test quality control in a computer-assisted semen analysis system. *Fertil Steril*. 2000;73(3):636–40. [https://doi.org/10.1016/S0015-0282\(99\)00575-0](https://doi.org/10.1016/S0015-0282(99)00575-0).
9. Bossé MJ, Lee TD, Swinson M, Faulconer J. The NCTM Process Standards and the Five < i > Es of Science: Connecting Math and Science. *School Sci Math*. 2010. <https://doi.org/10.1111/j.1949-8594.2010.00033.x>.

10. Bybee RW, Ferrini-Mundy J, Loucks-Horsley S. National Standards and School Science and Mathematics. *School Sci Math*. 1997. <https://doi.org/10.1111/j.1949-8594.1997.tb17282.x>.
11. Chin H, Meng CC, Yew WT, Musa M. Validating the Cognitive Diagnostic Assessment and Assessing Students' Mastery of 'Parallel and Perpendicular Lines' Using the Rasch Model. *Participatory Educational Res*. 2022. <https://doi.org/10.17275/per.22.147.9.6>.
12. Colonnese MW. The Development of Instructional Guidelines for Elementary Mathematical Writing. *School Sci Math*. 2020. <https://doi.org/10.1111/ssm.12391>.
13. Deeba F, Ahmad I. A Rasch Analysis of Achievement Goal Orientation Scale in Language Education Context. *Nust J Social Sci Humanities*. 2023. <https://doi.org/10.51732/njssh.v9i2.171>.
14. Dwiliesanti WG, Yudiarso A. (2022). Rasch Analysis of the Indonesian Version of Individual Work Performance Questionnaire (IWPQ). *Jp3i (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*. <https://doi.org/10.15408/jp3i.v11i2.24157>
15. Eddy CM, Fuentes SQ, Ward EK, Parker YA, Cooper S, Jasper WA, Mallam W, Sorto MA, Wilkerson TL. Unifying the Algebra for All Movement. *J Adv Acad*. 2015. <https://doi.org/10.1177/1932202x14562393>.
16. Embretson SE, Reise SP. Item Response Theory for Psychologists. *Psychology*; 2013. <https://doi.org/10.4324/9781410605269>.
17. Eslahi Z, Alimoradi Z, Bahrami N, Lin C, Griffiths MD, Pakpour AH. Psychometric Properties of Postpartum Partner Support Scale—Persian Version. *Nurs Open*. 2021. <https://doi.org/10.1002/nop2.806>.
18. Fan J, Knoch U. (2019). Fairness in Language Assessment: What Can the Rasch Model Offer? *Studies in Language Assessment*. <https://doi.org/10.58379/jrwg5233>
19. Fan X. Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educ Psychol Meas*. 1998;58(3):357–81. <https://doi.org/10.1177/0013164498058003001>.
20. Fritz-Stratmann A, Balzer L, Herholdt R, Ragpot L, Ehler A. A Mathematics Competence Test for Grade 1 Children Migrates From Germany to South Africa. *South Afr J Child Educ*. 2014. <https://doi.org/10.4102/sajce.v4i2.207>.
21. Hammond A, Tennant A, Ching A, Parker J, Prior Y, Gignac M, Verstappen MA, S., and, O'Brien R. Psychometric Testing of the British-English Long-Term Conditions Job Strain Scale, Long-Term Conditions Work Spillover Scale and Work-Health-Personal Life Perceptions Scale in Four Rheumatic and Musculoskeletal Conditions. *Musculoskelet Care*. 2023. <https://doi.org/10.1002/msc.1774>.
22. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1–73. <https://doi.org/10.1111/jedm.12000>.
23. de Klerk S. Assessment of Structural and Cross-Cultural Validity of the Disabilities of the Arm, Shoulder and Hand Questionnaire: A Scoping Review. *Hand Therapy*. 2022. <https://doi.org/10.1177/17589983221140433>.

24. Koirala HP, Bowman JK. Preparing Middle Level Preservice Teachers to Integrate Mathematics and Science: Problems and Possibilities. *School Sci Math*. 2003. <https://doi.org/10.1111/j.1949-8594.2003.tb18231.x>.
25. Malonisio MO, Malonisio CC. Validation of the Teacher Education Institution's Entrance Test Using the Rasch Model. *Int J Innovative Res Sci Stud*. 2023. <https://doi.org/10.53894/ijirss.v6i3.1726>.
26. Manz PH, Gernhart AL, Bracaliello CB, Pressimone VJ, Eisenberg R. Preliminary Development of the Parent Involvement in Early Learning Scale for Low-Income Families Enrolled in a Child-Development-Focused Home Visiting Program. *J Early Intervention*. 2014. <https://doi.org/10.1177/1053815115573077>.
27. Messick S. Validity. *Educational measurement*. 3rd ed. American Council on Education; 1989. pp. 13–103.
28. Mohamed RH, Khalil I, Awaji B. Mathematics Teachers' Awareness of Effective Teaching Practices: A Comparative Study. *Eurasia J Math Sci Technol Educ*. 2023. <https://doi.org/10.29333/ejmste/12962>.
29. National Council of Teachers of Mathematics. *Principles and Standards for School Mathematics*. NCTM; 2000.
30. National Research Council. *Adding It Up: Helping Children Learn Mathematics*. National Academies; 2001. <https://doi.org/10.17226/9822>.
31. Nick JM, Durkin AC, Caroci-Becker A, de Almeida Cruz D, Zorzim VI, Wonder AH. Validation of the Brazilian Portuguese Version of Evidence-Based Practice Knowledge Assessment in Nursing (EKAN). *Revista Da Escola De Enfermagem Da Usp*. 2021. <https://doi.org/10.1590/s1980-220x2020012003768>.
32. Novak E, Soytürk İ, Navy SL. Development of the Science Teaching Anxiety Scale for Preservice Elementary Teachers: A Rasch Analysis. *Sci Educ*. 2022. <https://doi.org/10.1002/sce.21707>.
33. Pellegrino JL, Charlton NP, Goolsby C. (2020). Stop the Bleed Education Assessment Tool (SBEAT): Development and Validation. *Cureus*. <https://doi.org/10.7759/cureus.10567>
34. Pellegrino JW, Chudowsky N, Glaser R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*.
35. Prior Y, Tennant A, Tyson S, Kjekken I, Hammond A. Measure of Activity Performance of the Hand (MAP-Hand) Questionnaire: Linguistic Validation, Cultural Adaptation and Psychometric Testing in People With Rheumatoid Arthritis in the UK. *BMC Musculoskelet Disord*. 2018. <https://doi.org/10.1186/s12891-018-2177-5>.
36. Putri MRN. Major Challenges of Teachers Implementing Nature-Based Education Concept Secondary School Students In Indonesia. *J Social Political Sci Soc*. 2023;2(1). Article 1. <https://doi.org/10.58330/society.v2i1.266>.
37. Qudratuddarsi H, Hidayat R, Shah Z, Nasir RL, Imami NW, M. K., and, Nor M. R. bin. (2022). Rasch Validation of Instrument Measuring Gen-Z Science, Technology, Engineering, and Mathematics

- (STEM) Application in Teaching During the Pandemic. *International Journal of Learning Teaching and Educational Research*. <https://doi.org/10.26803/ijlter.21.6.7>
38. Raman K, Siegert RJ, Bharatharaj J, Krägeloh CU. Validation of a Tamil Version of the Five Facet Mindfulness Questionnaire Using Rasch Analysis. *Indian J Psychol Med*. 2024. <https://doi.org/10.1177/02537176231220564>.
 39. Segar TN. The Psychometric Assessment of GRIT Using Rasch Measurement Model for Vernacular Students Setting. *Semarak Int J Appl Psychol*. 2024. <https://doi.org/10.37934/sijap.2.1.4457>.
 40. Sorrentino G, Vercelli S, Salgovic L, Ronconi G, Bakhsh HR, Ferriero G. Psychometric Properties of the Client Satisfaction With Device Module of the Orthotics and Prosthetics Users' Survey (OPUS): A Scoping Review. *Int J Rehabil Res*. 2021. <https://doi.org/10.1097/mrr.0000000000000484>.
 41. Sorto MA. Data Analysis and Statistics in Middle Grades: An Analysis of Content Standards. *School Sci Math*. 2011. <https://doi.org/10.1111/j.1949-8594.2010.00068.x>.
 42. Suryadi A, Fawaiz S, Kurniati E, Swandi A. Gender- And Academic Level-Bias in MATS When Measuring Attitude Towards Science in Indonesia: A Rasch Analysis. *J Pedagogical Res*. 2023. <https://doi.org/10.33902/jpr.202422983>.
 43. Syawaludin A, Supriyati Y, Rahayu W. (2019). *RASCH Model Application for Validation of Measurement Instruments of Student Nationalism*. <https://doi.org/10.17501/24246700.2019.5204>
 44. Veith JM, Bitzenbauer P, Girnat B. Assessing Learners' Conceptual Understanding of Introductory Group Theory Using the CI2GT: Development and Analysis of a Concept Inventory. *Educ Sci*. 2022. <https://doi.org/10.3390/educsci12060376>.
 45. Wijaya T. (2023). *Unpacking the Fintech Regulatory Sandbox Framework in Indonesia: Risks Management and the Data Privacy Imperative*. <https://doi.org/10.35497/565201>
 46. Wong CM, Hassan MR, Hod R, Wan Puteh SE, Bakar SA. A Tale of Dual-Approach Construct Validation and Reliability Testing for a < i> Zika Infection Awareness Knowledge Questionnaire. *J Health Res*. 2021. <https://doi.org/10.1108/jhr-08-2020-0332>.

Figures

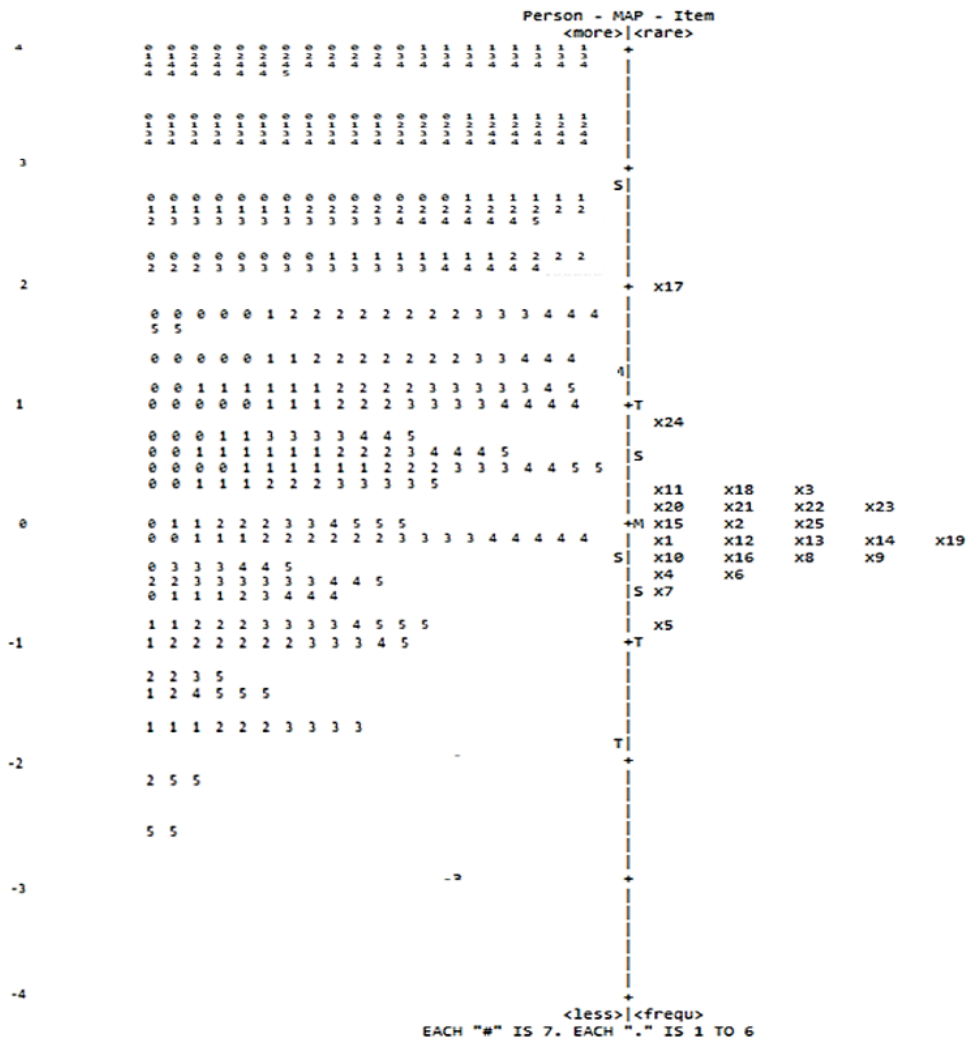


Figure 1

Wright map showing the distribution of student ability and item difficulty.

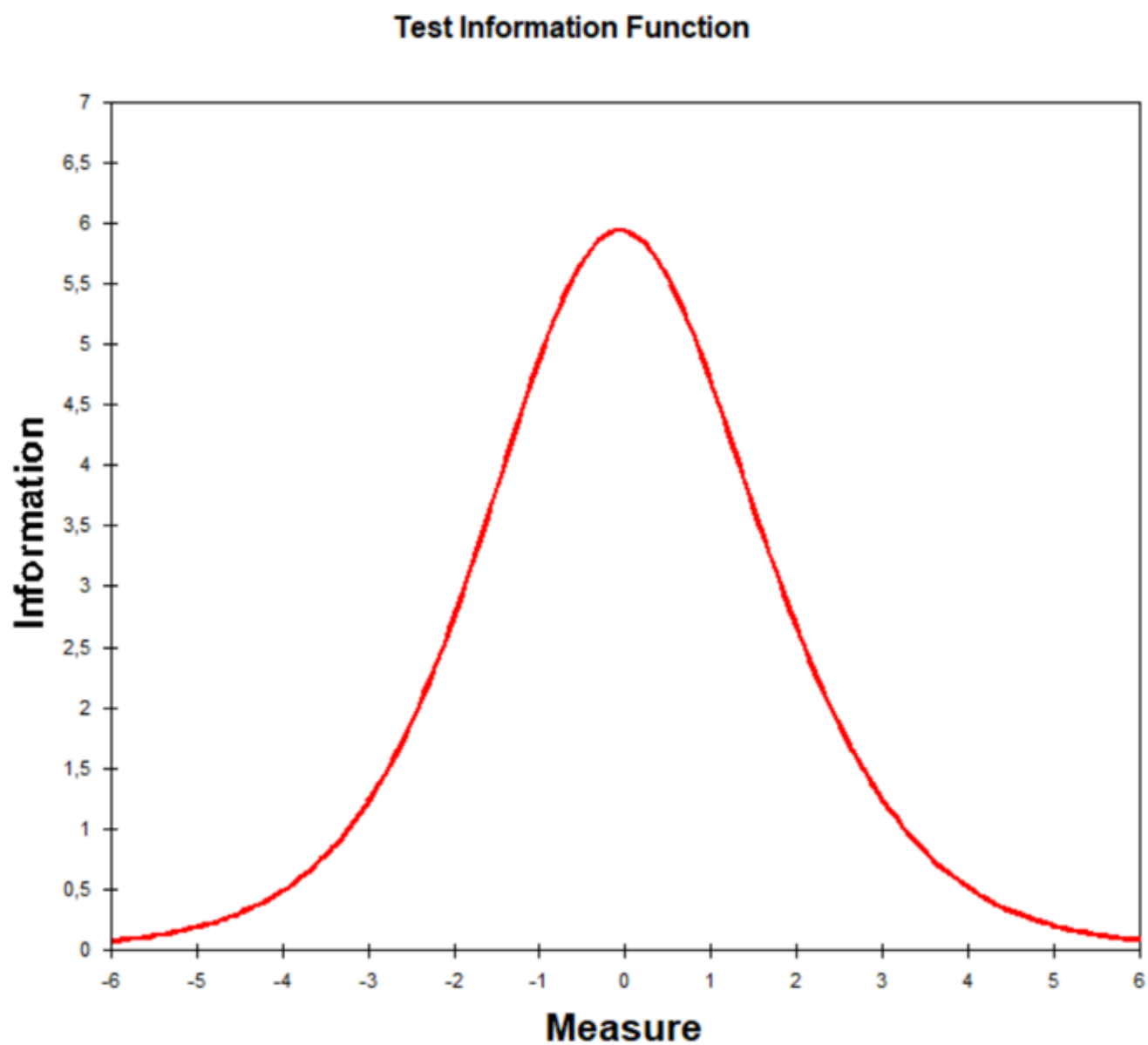


Figure 2

Test information function of the mathematical communication skills instrument.