Validating the Mathematical Literacy Test using the Rasch Measurement Model

Shahibul Ahyan[1,*], Turmudi Turmudi[2], and Dadang Juandi[2]

[1] Universitas Hamzanwadi, East Lombok,West Nusa Tenggara, 83612, Indonesia
[2] Universitas Pendidikan Indonesia, Bandung, West Java, 40154, Indonesia
*shahibulahyan@hamzanwadi.ac.id

Abstract. Mathematical literacy is an essential competency that enables students to apply mathematical understanding and reasoning to address everyday challenges. This research sought to develop and validate a set of mathematical literacy items through a comprehensive psychometric approach grounded in Rasch measurement theory. The study involved three experts for content validation and 275 junior high school students, of whom 212 provided data included in the empirical Rasch analysis. Expert judgment data were analyzed using Many-Facet Rasch Measurement with FACETS version 3.83.6 to assess content validity while accounting for rater severity. Student response data were evaluated using the Rasch model with Winsteps version 4.8.0.0 to examine item performance, dimensionality, differential item functioning (DIF) by gender, and reliability. The results indicated that 10 of the 14 initial items met the content-validation criteria based on expert judgment. Rasch analysis of student responses showed that the 10 retained items generally functioned as intended, although IN1 and EM3 exhibited overfit and FO1 showed statistically significant gender-related DIF. IN1 and EM3 were retained because overfit alone was not considered sufficient evidence for item removal, while FO1 was identified for further review. The instrument showed high item reliability and separation, whereas person reliability and separation were relatively low, indicating limited precision in distinguishing individual ability levels. Overall, the findings provide evidence supporting the use of the 10-item instrument for group-level assessment of mathematical literacy, while further refinement and validation are needed to improve person-level measurement precision.

Keywords: Mathematical Literacy, Rasch Measurement, Many-Facet Rasch Measurement, Junior High School Students.

1 Introduction

Mathematical literacy has become a significant concern in many countries due to its importance in modern life and work. Employers are worried that students leave school without the necessary mathematical skills for the workplace [1]. The ability to interpret contextual problems into mathematical language is crucial, yet many countries, including Indonesia, still struggle with low mathematical literacy skills [2, 3].

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_12

Interestingly, there are contradictions in how mathematical literacy is approached. While governments have included contextual aspects in school curricula, many schools still lack the ability to teach contextual skills effectively [2]. Additionally, specific attempts to develop mathematical literacy through mathematics alone have not been successful, and teaching for mathematical literacy appears to require non-traditional methods [4].

The development of mathematical literacy in schools remains a challenge. Research suggests that both researchers and teachers are uncertain about how to develop students' mathematical literacy effectively [4]. The concept of mathematical proficiency has been proposed as a focus for both research and practice to avoid extreme positions in disputes over school mathematics [5]. Moving forward, more qualitative research is needed, particularly in classroom studies focusing on teachers' priorities and best practices for developing mathematical literacy [4].

Addressing mathematical literacy issues requires a comprehensive approach. This includes developing PISA-like tasks to improve students' mathematical literacy [6, 7], considering the role of home environments in numeracy development [8, 9], and recognizing the complex relationship between mathematical knowledge and literacy [10]. Efforts to enhance mathematical literacy must consider these various factors to effectively improve students' skills and prepare them for real-world challenges.

Developing effective mathematical literacy tests requires a multifaceted approach that considers various aspects of mathematical proficiency, information processing, and real-world applications. Such tests can help identify areas for improvement and guide the development of targeted interventions to enhance students' mathematical literacy skills [11, 12]. By incorporating elements from realistic mathematics education and problem-solving processes, these tests can provide a more comprehensive assessment of students' mathematical abilities and their readiness to apply mathematical knowledge in real-life situations.

To ensure that the mathematical literacy test can assess the students’ mathematical literacy, the test should valid and reliable. One of the models to measure the test is using Rasch measurement. Rasch measurement models offer several advantages for validating mathematical literacy tests. They can assess unidimensionality, item fit, response thresholds, targeting, and internal consistency [13, 14]. Additionally, Rasch analysis allows for the transformation of ordinal data into interval-level measurements, enhancing the accuracy and suitability of data for parametric statistical tests [15]. These features make Rasch analysis a valuable tool for developing and validating robust mathematical literacy assessment. Therefore, this study aims to validate the mathematical literacy test using Rasch measurement.

2 Method

This study employed a cross-sectional research design. Cross-sectional research design is perhaps the most common design in the social sciences, occurring when researchers collect data from a group of research participants at a single point in time using instruments, such as tests, questionnaires, interviews, or observations [16, 17]. A cross-sectional design was appropriate because the data were collected from the participants at a single point in time [18]. In addition, cross-sectional research helps researchers compare several variables at the same time.

The mathematical literacy test used in this study is a mathematical literacy test that follows the framework of the PISA mathematics questions published by the OECD. The initial instrument consisted of 14 mathematical literacy items developed in accordance with the PISA mathematics framework. The items represented three mathematical processes: four items targeting formulating situations mathematically, five items targeting employing mathematical concepts, procedures, facts, and reasoning, and five items targeting interpreting and evaluating mathematical results. Table 1 presents the distribution of the items across mathematical literacy processes, content domains, and indicators.

Table 1. The process, content and indicator of mathematical literacy test.

Mathematical Literacy Process	Content	Indicator	No.	Question Code
Formulate the situation mathematically	Change and relationships	Identify mathematical aspects of a problem situated in a real-world context and identify significant variables.	2, 3	ML2, ML3
	Space and shape	Recognize mathematical structures (including regularities, relationships and patterns) in problems or situations	4, 5	ML4, ML5
Employ concepts, procedures, facts, and reasoning	Change and relationships	Apply mathematical facts, rules, algorithms, and structures when finding solutions	1	ML1
	Quantity	Design and apply strategies to find mathematical solutions	7, 8, 9	ML7, ML8, ML9
	Space and shape	Design and apply strategies to find mathematical solutions	11	ML11
Interpreting and evaluating mathematical results	Change and relationships	Reinterpret mathematical results into a real-world context	6	ML6

Mathematical Literacy Process	Content	Indicator	No.	Question Code
	Quantity	Evaluate the reasonableness of mathematical solutions in the context of real-world problems	12	ML12
	Space and shape	Reinterpret mathematical results into a real-world context	10	ML10
	Uncertainty and data	Evaluate the reasonableness of mathematical solutions in the context of real-world problems	13, 14	ML13, ML14

This study involved three experts and 275 junior high school students. The three experts participated in the content-validation stage, while student responses were used for empirical construct validation. Of the 275 students who participated in the study, data from 212 students were included in the Rasch analysis after data screening, while the remaining responses were excluded because they did not meet the criteria for analysis.

The results of the expert review were analyzed using Many-Facets Rasch Measurement (MFRM) with the help of the Facets version 3.83.6 application. MFRM is one of the analysis models that is a development of the Rasch Model [19]. The MFRM model can include more than two variables/facets in the analysis, which makes it very suitable for performance assessment that includes several facets such as examinees, assessors, assessment criteria, and tasks [20, 21]. In this study, there are three facets or variables analyzed using Facets software, namely experts, items, and criteria. For mathematical literacy questions, there are three experts: Rater 1, Rater 2, and Rater 3; there are 14 items; and there are five criteria, namely ability, process, content, context, and sentence structure. For the choice of criteria, experts rated from strongly disagree (1) to strongly agree (5). The indicators used in assessing the results of experts using MFRM [22], which fulfills at least two of the following three indicators:

- Outfit mean square (MNSQ) value: $0.5 < \text{MNSQ} < 1.5$
- Z-standard (ZSTD) Outfit value: $-2 < \text{ZSTD} < 2$
- Point Measure Correlation (Pt Mean Corr) value: $0.4 < \text{Pt Mean Corr} < 0.85$

Data from 212 students were analyzed using Rasch Model analysis with the help of Winsteps software version 4.8.0.0. There are four analyses that will be discussed from this Rasch Model, namely dimensionality analysis, item analysis, and reliability analysis.

Dimensional analysis was conducted to determine whether all items in the mathematical literacy questionnaire were able to measure one variable regarding

mathematical literacy. This unidimensionality is a fundamental requirement of the Rasch Model. The criteria for unidimensionality are if the raw variance explained by measure is at least 20%. If the value is more than 40%, it means good, and if more than 60%, it means excellent [23]. Furthermore, item functioning was examined using item difficulty estimates, fit statistics, and point-measure correlations. Infit and outfit mean-square statistics were used to identify response patterns that deviated from Rasch model expectations. Items falling outside the adopted fit range were examined further rather than automatically removed, as decisions regarding item retention also considered the direction and magnitude of misfit and the item's relevance to the intended mathematical literacy construct [23, 26]. Differential Item Functioning (DIF) was also examined to determine whether individual items functioned differently across gender groups after accounting for differences in mathematical literacy ability. Based on the Mantel–Haenszel procedure, items with probability values below 0.05 were identified as showing statistically significant DIF and were subjected to further substantive review rather than being automatically classified as biased or removed [23]. Finally, measurement reliability was evaluated using person reliability, item reliability, person and item separation indices, and Cronbach's alpha (KR-20).

3 Results and Discussion

3.1 Perceived Item Quality Analysis based on Expert Judgment Ratings

Fourteen mathematical literacy items (ML1–ML14) were evaluated by three experts using five criteria: ability, process, content, context, and sentence structure. Each criterion was rated on a five-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). The ratings were analyzed using the Many-Facet Rasch Measurement (MFRM) model with FACETS version 3.83.6.

Figure 1 below displays a map of variables representing expert calibration, items, and criteria with five rating scales used by raters to assess mathematical literacy questions. This map is also called the “Wright map”.

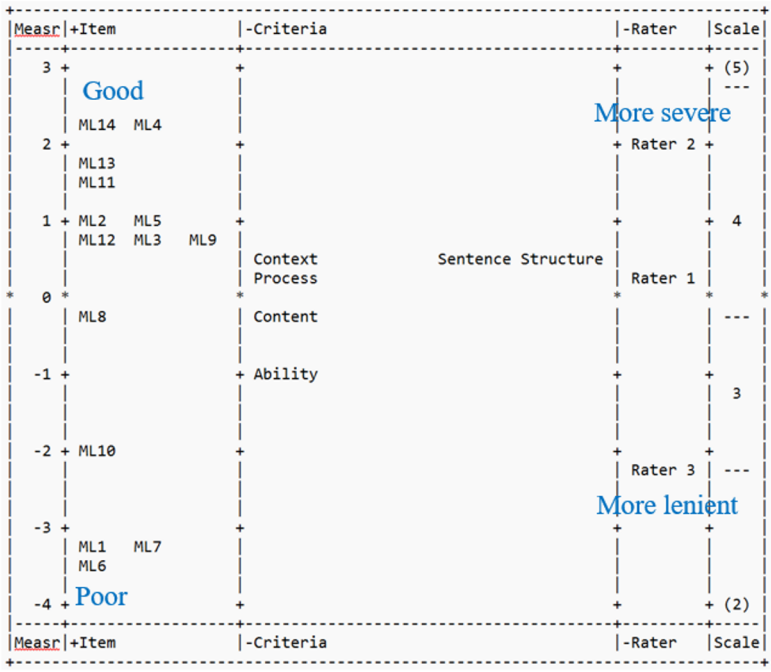


Fig 1. Wright map of mathematical literacy test

Figure 1 presents the variable map from the MFRM analysis of expert ratings, showing the relative locations of items, evaluation criteria, raters, and rating-scale categories on the common logit scale. The first column presents the measurement scale in logit units, followed by items related to mathematical literacy, the five assessment criteria, the expert raters, and the rating categories. The majority of items are positioned above 0 logits, indicating generally favorable evaluations from experts. However, items ML6, ML1, and ML7 are situated between −4 and −3 logits, suggesting these items received relatively low expert support and struggled to achieve ratings above scale point 3, which signifies moderate agreement. In contrast, items ML14 and ML4 are located at the higher end of the logit scale, indicating strong expert consensus and identifying them as the highest-quality items. Among the five criteria, ability occupied the lowest location on the criterion facet, indicating that items tended to receive relatively more favorable ratings on this criterion. Rater 3 has the lowest severity measure, confirming this rater as the most lenient.

Content validation was conducted through expert judgment to examine the quality and appropriateness of the mathematical literacy items before empirical testing with students. The experts evaluated each item using five criteria: ability, process, content, context, and sentence structure. These criteria were intended to capture complementary aspects of item quality, ranging from alignment with the intended mathematical literacy competency to the clarity of the language used in the item.

The ability criterion refers to the extent to which an item represents the intended mathematical literacy competency and elicits the type of mathematical reasoning targeted by the assessment. The process criterion concerns whether the item appropriately engages the mathematical literacy processes of formulating situations mathematically, employing mathematical concepts and procedures, and interpreting or evaluating mathematical results. The content criterion addresses the relevance and mathematical accuracy of the concepts, procedures, and representations embedded in the item. The context criterion examines whether the problem is situated in a meaningful and appropriate real-world situation and whether the context contributes substantively to the mathematical task. Finally, the sentence structure criterion concerns the linguistic quality of the item, including grammatical accuracy, clarity, readability, comprehensibility, and the absence of wording that may introduce unintended ambiguity.

The operational definitions of the five criteria are presented in Table 2.

Table 2. Operational definitions of the expert-judgment criteria

Criterion	Operational Definition	Focus of Judgment
Ability	The extent to which the item represents the intended mathematical literacy competency and elicits the targeted type of mathematical reasoning.	Alignment with the intended mathematical literacy competency
Process	The extent to which the item engages the intended mathematical literacy processes of formulating, employing, and interpreting or evaluating mathematics.	Appropriateness of the mathematical literacy process
Content	The extent to which the mathematical concepts, procedures, and representations embedded in the item are relevant and mathematically accurate.	Relevance and mathematical accuracy
Context	The extent to which the item is situated in a meaningful and appropriate real-world situation that contributes substantively to the mathematical task.	Relevance and meaningfulness of the context
Sentence Structure	The extent to which the item is grammatically appropriate, clear, readable, comprehensible, and free from unintended linguistic ambiguity	Linguistic clarity and comprehensibility

Each item was rated on all five criteria using a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scale represented the extent to which an item satisfied each criterion. A rating of 1 indicated very weak conformity with the criterion, 2 indicated limited conformity, 3 indicated moderate or uncertain conformity, 4 indicated good conformity, and 5 indicated very strong conformity. Thus, higher observed ratings represented more favorable expert judgments regarding the extent to which an item met the specified criterion.

The expert ratings were analyzed using the Many-Facet Rasch Measurement (MFRM) model to examine the quality of the mathematical literacy items while accounting for variation in the rating process. The model enables item evaluations to be calibrated while considering differences in rater severity or leniency and the functioning of the rating categories [24, 25]. This was important because experts may differ in how strictly they apply the same evaluation criteria, and such differences may influence the ratings assigned to individual items.

The distribution of item quality measures derived from the expert ratings is presented in Table 3. The variation in item locations indicates that the items differed in the extent to which they were judged to meet the predefined criteria of ability, process, content, context, and sentence structure. Accordingly, these measures reflect differences in expert-judged item quality, rather than differences in empirical item difficulty or in the level of student ability required to answer the items [24, 25]. Items receiving more favorable expert evaluations showed stronger conformity with the intended characteristics of the mathematical literacy assessment, whereas less favorable evaluations indicated aspects that warranted closer examination and possible revision. Empirical item difficulty was examined separately in the subsequent Rasch analysis using student response data.

Table 3. Expert measurement report

Rater	Severity Measure	SE	Infit	Oufit	Fair (M) Average	Obs. Average	Number of Rating
Rater 2	2.00	0.20	1.19	0.88	2.63	2.90	203
Rater 1	0.16	0.21	0.88	0.82	3.58	3.51	249
Rater 3	-2.27	0.22	0.70	0.98	4.39	4.27	299
Average	-0.03	0.21	0.92	0.89	3.54	3.56	249.3

The rater measurement results from the Many-Facet Rasch Measurement (MFRM) analysis are presented in Table 3. The raters differed in severity, with Rater 2 being the most severe (2.00 logits), Rater 3 the most lenient (−2.27 logits), and Rater 1 positioned close to the average (0.16 logits). Despite these differences, all raters showed acceptable fit to the MFRM model, with infit and outfit mean-square values within the recommended range of 0.5–1.5 [26, 27], while the relatively small standard errors (0.20–0.22) indicated adequate precision of the severity estimates [28, 29]. The item-level results based on the expert ratings are presented in Table 4.

Table 4. MFRM item quality measures based on expert ratings

Item	Measure	SE	Total Score
ML4	2.33	0.5	65
ML14	2.33	0.5	65
ML13	1.64	0.46	62
ML11	1.44	0.45	61
ML2	1.05	0.43	59
ML5	1.05	0.43	59
ML9	0.87	0.43	58
ML12	0.87	0.43	58
ML3	0.69	0.42	57
ML8	-0.17	0.41	52
ML10	-2.05	0.47	42
ML1	-3.25	0.52	37
ML7	-3.25	0.52	37
ML6	-3.53	0.54	36
Average	0.00	0.47	53.4

Table 4 presents the MFRM item measures derived from the expert ratings. Because the experts evaluated each item against the five criteria of ability, process, content, context, and sentence structure, the resulting measures represent the relative quality of the items as judged against these criteria, rather than their empirical difficulty for students [26, 27]. In the parameterization used in this study, a higher item measure indicates a more favorable overall expert evaluation and, consequently, stronger conformity with the predefined quality criteria. Conversely, a lower measure indicates a less favorable evaluation and weaker conformity with these criteria. Thus, the item locations are interpreted along a continuum of expert-judged item quality, rather than in terms of how easy or difficult the items are for students to answer.

As shown in Table 4, the items varied in their relative positions on this expert-judgment continuum. ML4 and ML14 had the highest measures (2.33 logits) and total scores (65), indicating the most favorable overall evaluations across the five criteria. ML13 (1.64 logits; total score = 62) and ML11 (1.44 logits; total score = 61) also received relatively favorable evaluations. At the other end of the continuum, ML1 and ML7 had measures of −3.25 logits (total score = 37), while ML6 had the lowest measure (−3.53 logits) and total score (36). These lower measures should not be interpreted as indicating easier items; rather, they indicate that the items showed comparatively weaker conformity with the predefined quality criteria based on the experts' evaluations.

Accordingly, the MFRM item measures were used to characterize differences in the relative quality of the items within the expert-judgment framework and to support the content-validation process. Empirical item difficulty and psychometric functioning were evaluated separately using student response data [26, 27, 30].

3.2 Item Difficulty and Fit Analysis based on Student Responses

In order to evaluate whether the mathematical literacy tool assesses a singular underlying construct, a dimensionality analysis was performed in alignment with the principles of Rasch measurement theory. The Rasch model inherently presupposes unidimensionality, suggesting that all items collectively evaluate one primary latent trait, which, in this context, is mathematical literacy. According to international Rasch measurement guidelines, it is generally accepted that the raw variance explained by Rasch measures should exceed 20%, and any unexplained contrasts should not indicate significant secondary dimensions [26, 31].

Table 5. Standardized residual variance (in Eigenvalue units)

		Eigenvalue	Observed		Expected
Total raw variance in observations	=	17.4624	100.0%		100.0%
Raw variance explained by measures	=	7.4624	**42.7%**		49.1%
Raw unexplained variance (total)	=	10.0000	57.3%	100.0%	50.9%
Unexplned variance in 1st contrast	=	1.9432	11.1%	19.4%	
Unexplned variance in 2nd contrast	=	1.5038	8.6%	15.0%	
Unexplned variance in 3rd contrast	=	1.2467	7.1%	12.5%	
Unexplned variance in 4th contrast	=	1.1090	6.4%	11.1%	
Unexplned variance in 5th contrast	=	1.0129	5.8%	10.1%	

Table 5 indicates that the Rasch measures account for 42.7% of the total observed variance, surpassing the widely accepted threshold of 20% and strongly suggesting unidimensionality. According to Linacre, this level of explained variance is adequate to confirm the presence of a primary latent construct in educational measurement [32]. The Rasch measures explained 42.7% of the observed variance. The first residual contrast had an eigenvalue of 1.94, below the commonly used benchmark of 2.0, providing reasonable support for the absence of a substantial secondary dimension [26, 32]. Taken together, these results support the interpretation of the instrument as measuring a dominant mathematical literacy construct. Further analysis using principal component analysis of standardized residuals supports this interpretation, as the first

contrast yields an eigenvalue of 1.94, which is below the critical threshold of 2.0, indicating no significant secondary dimension [26]. Subsequent contrasts exhibit even smaller eigenvalues, further affirming the absence of significant alternative constructs. Overall, these findings demonstrate that the items, despite representing various mathematical processes, consistently measure a single mathematical literacy construct, consistent with international assessment frameworks such as those of the OECD [33].

The empirical evaluation of the items focused on item difficulty, fit to the Rasch model, and differential item functioning (DIF). Item difficulty was interpreted descriptively to examine the distribution and targeting of the items, whereas decisions regarding item functioning were informed primarily by fit statistics and DIF evidence.

Item Measure

In Rasch analysis, item measures denote the difficulty level of each mathematical literacy question, indicating the likelihood of respondents answering correctly. Items with higher logit values are more challenging, whereas those with lower (negative) logit values are easier. Figure 2 displays the item measure outcomes derived from the Rasch analysis conducted on responses from 212 students. Within the framework of Rasch measurement theory, representing item difficulty in logits situates them on a uniform interval scale, enabling direct comparison between item difficulty and individual ability. This characteristic is a notable advantage of Rasch modeling over classical test theory, facilitating a more precise interpretation of item performance [26].

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
8	25	212	1.30	.19	.70	-1.19	.44	-2.15	.42	.25	90.6	89.2	IN1
7	30	212	1.13	.18	.62	-1.74	.45	-2.23	.44	.28	87.7	87.3	EM3
4	56	212	.56	.13	.57	-2.70	.51	-2.50	.53	.37	78.3	76.8	FO4
3	86	212	.17	.10	.73	-1.90	.61	-2.22	.56	.44	71.2	68.0	FO3
9	92	212	.10	.10	1.58	3.35	.73	-1.48	.54	.45	73.6	64.6	IN2
5	112	212	-.08	.09	.94	-.42	.91	-.48	.52	.48	62.7	61.2	EM1
2	171	212	-.50	.08	1.14	1.17	1.10	.72	.53	.56	50.0	52.0	FO2
10	194	212	-.63	.07	1.05	.47	1.40	2.53	.42	.58	36.3	47.8	IN3
6	195	212	-.64	.07	1.11	.99	.94	-.37	.66	.58	44.8	47.8	EM2
1	370	212	-1.41	.06	1.48	4.32	1.37	3.12	.51	.69	27.8	34.0	FO1
MEAN	133.1	212.0	.00	.11	.99	.2	.85	-.5			62.3	62.9	
P.SD	98.8	.0	.80	.04	.33	2.2	.34	1.9			20.6	17.1	

Fig. 2. Item measure of 10 mathematical literacy items

The item measure results indicate a significant and extensive range of difficulty, spanning from 1.30 logits (IN1) to −1.41 logits (FO1). Item IN1 is identified as the most challenging, suggesting that a relatively high level of mathematical literacy is

necessary for successful completion. Experts in international measurement underscore the importance of such difficult items in distinguishing students with higher abilities and accurately representing the upper spectrum of the construct, thereby minimizing ceiling effects and enhancing measurement accuracy [28]. Conversely, item FO1 is the simplest, indicating its accessibility to the majority of students. In Rasch-based assessment design, easy items are crucial for anchoring the lower end of the ability spectrum and ensuring fair classification of students with lower performance [29].

Item difficulty varied across the processes of formulating, employing, and interpreting, with some interpreting items located at relatively higher difficulty levels. In this set of items, tasks that required students to interpret mathematical results and relate them to real-world contexts tended to be more challenging. However, this pattern should be interpreted with caution given the limited number of items representing each process. It does not necessarily suggest that interpreting is inherently more difficult than formulating or employing, as item difficulty may also reflect differences in content, context, and the specific demands of each task within the mathematical literacy framework [30]. The associated standard errors are small (approximately 0.06–0.19), indicating stable and precise estimation of item parameters. Overall, the broad and well-targeted distribution of item difficulty supports a robust representation of the construct and enhances the validity and interpretability of the mathematical literacy instrument [26].

Fit Statistics

Figure 3 delineates the item fit statistics derived from Rasch measurement analysis, which are crucial for assessing the degree to which individual items align with model expectations. An appropriate item fit signifies that an item substantially contributes to the measurement of a singular latent construct and operates consistently across respondents with varying ability levels [26, 28]. The evaluation of item fit was conducted using three established Rasch criteria: outfit mean square values ranging from 0.5 to 1.5, outfit Z-standardized values between −2 and +2, and point–measure correlations from 0.40 to 0.85. In alignment with international guidelines, conclusions regarding item misfit were drawn based on the convergence of multiple indicators rather than reliance on a single statistic [31].

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
9	92	212	.10	.10	1.58	3.35	.73	-1.48	A .54	.45	73.6	64.6	IN2
1	370	212	-1.41	.06	1.48	4.32	1.37	3.12	B .51	.69	27.8	34.0	FO1
10	194	212	-.63	.07	1.05	.47	1.40	2.53	C .42	.58	36.3	47.8	IN3
2	171	212	-.50	.08	1.14	1.17	1.10	.72	D .53	.56	50.0	52.0	FO2
6	195	212	-.64	.07	1.11	.99	.94	-.37	E .66	.58	44.8	47.8	EM2
5	112	212	-.08	.09	.94	-.42	.91	-.48	e .52	.48	62.7	61.2	EM1
3	86	212	.17	.10	.73	-1.90	.61	-2.22	d .56	.44	71.2	68.0	FO3
8	25	212	1.30	.19	.70	-1.19	.44	-2.15	c .42	.25	90.6	89.2	IN1
7	30	212	1.13	.18	.62	-1.74	.45	-2.23	b .44	.28	87.7	87.3	EM3
4	56	212	.56	.13	.57	-2.70	.51	-2.50	a .53	.37	78.3	76.8	FO4
MEAN	133.1	212.0	.00	.11	.99	.2	.85	-.5			62.3	62.9	
P.SD	98.8	.0	.80	.04	.33	2.2	.34	1.9			20.6	17.1	

Fig. 3. Fit statistics of 10 mathematical literacy items

The findings demonstrate that the majority of items exhibit strong alignment with the Rasch model, as indicated by outfit mean square (MNSQ) values approximating 1.0, Z-standardized (ZSTD) values within acceptable parameters, and positive, moderate point–measure correlations. These results imply that most items function as predicted by the model and are consistent with the concept of mathematical literacy. According to Wright, well-fitting items avoid both excessive randomness (underfit) and excessive predictability (overfit), thereby facilitating effective measurement along the latent continuum [28].

IN1 and EM3 showed outfit MNSQ values of 0.44 and 0.45, respectively, with corresponding ZSTD values of −2.15 and −2.23. These values fall slightly below the adopted MNSQ range and indicate overfit, meaning that responses to these items were more predictable than expected by the Rasch model. Their point-measure correlations remained positive (IN1 = 0.42; EM3 = 0.44), indicating that both items were positively associated with the underlying mathematical literacy measure.

The overfit observed for IN1 (interpreting) and EM3 (employing) was not considered sufficient evidence for item removal. Overfit indicates greater response predictability than expected by the model, but does not by itself demonstrate that an item fails to represent the intended construct. Given their continued relevance to the mathematical literacy processes represented in the instrument, IN1 and EM3 were retained. Their functioning should nevertheless be examined further in future studies with larger and more diverse samples [26, 27, 30].

Differential Item Functioning (DIF)

To evaluate the effectiveness of mathematical literacy items for both male and female students, a Differential Item Functioning (DIF) analysis was performed. DIF analysis is essential for construct validity, as it determines whether observed differences between groups arise from genuine variations in the latent trait or result from item bias [26]. The Mantel–Haenszel chi-square statistic, a widely recognized method in Rasch-based and large-scale assessments, was utilized to assess uniform DIF. In accordance with international standards, items with probability values below 0.05 were identified as having statistically significant DIF, ensuring a balance between detecting bias and controlling false-positive results [34].

Table 6. DIF of 10 mathematical literacy items on gender

Person	Mantel		Name	Person	Mantel		Name
Class/	Chi-squ	Prob.		Class/	Chi-squ	Prob.	
L	106.411	0.0011	FO1	P	1064.11	0.0011	FO1
L	0.4029	0.5256	FO2	P	0.4029	0.5256	FO2
L	23.287	0.1270	FO3	P	23.287	0.1270	FO3
L	13.768	0.2406	FO4	P	13.768	0.2406	FO4
L	37.158	0.0539	EM1	P	37.158	0.0539	EM1
L	23.019	0.1292	EM2	P	23.019	0.1292	EM2
L	0.3333	0.5637	EM3	P	0.3333	0.5637	EM3
L	0.0667	0.7963	IN1	P	0.0667	0.7963	IN1
L	0.0415	0.8386	IN2	P	0.0415	0.8386	IN2
L	27.497	0.0973	IN3	P	27.497	0.0973	IN3

As shown in Table 6, most mathematical literacy items did not exhibit statistically significant differential item functioning (DIF) by gender. Specifically, FO2, FO3, FO4, EM1, EM2, EM3, IN1, IN2, and IN3 had probability values greater than 0.05, indicating no statistical evidence of gender-related DIF for these items. Thus, within the present sample, these items showed no significant differences in functioning between male and female students after accounting for their mathematical literacy ability. These findings provide evidence that most items functioned consistently across gender groups [26].

In contrast, FO1 exhibited statistically significant gender-related DIF, with a probability value of 0.0011, which is below the 0.05 threshold. This finding indicates that the response pattern for FO1 differed between male and female students after

accounting for differences in mathematical literacy ability. However, the presence of statistically significant DIF does not, by itself, establish that the item is biased. Rather, it signals the need for substantive examination of the item to determine whether construct-irrelevant factors, such as its context, wording, representation, or familiarity, may have contributed to the observed difference [34].

FO1 represents the formulating process, which requires students to translate a real-world situation into a mathematical representation. In context-based mathematical literacy assessments, item performance may be influenced not only by the mathematical demands of a task but also by students' familiarity with the situation in which the problem is presented. Previous research, including studies related to the PISA framework, has highlighted the importance of considering contextual characteristics when examining differential item functioning [35]. The DIF observed for FO1 should therefore be interpreted as evidence requiring further investigation rather than as direct evidence that the item is inherently biased.

Accordingly, FO1 was retained for substantive review rather than automatically removed from the instrument. Its context, wording, and representational features should be examined to identify possible sources of the observed gender-related DIF and revised where necessary. This approach preserves the representation of the formulating process in the instrument while allowing potential fairness concerns to be addressed. Further empirical testing of the revised item with a larger and more diverse sample is needed to determine whether the observed DIF persists.

3.3 Reliability Analysis

A reliability analysis was performed to assess the consistency and accuracy of the mathematical literacy instrument. In accordance with Rasch measurement theory, reliability was evaluated separately for individuals and items, underscoring the instrument's capacity to distinguish among individuals and the consistency of item difficulty estimates. Figure 4 presents the Winsteps results for 212 students and 10 items.

The person reliability coefficient identified in this study is 0.51, which is generally considered weak. Within the framework of Rasch measurement, person reliability reflects the instrument's ability to consistently differentiate students along the mathematical literacy scale. A low person reliability score is often associated with a narrow range of student abilities, a short test, or poor alignment between item difficulty and student ability [28]. In this study, the limited number of items (10 items) likely contributes to this, as shorter tests typically do not provide sufficient information to accurately distinguish between various proficiency levels.

This interpretation is further supported by the person separation index of 1.01, which falls below the generally accepted threshold of 2.0. Person separation measures how many statistically distinct ability levels an instrument can identify. A value near 1 implies that the instrument can only differentiate broad groups of students (e.g., higher and lower ability) rather than multiple, finely-tuned proficiency levels. As Linacre points out, low person separation does not necessarily indicate poor item quality but is often due to sample homogeneity or a short test length [32].

SUMMARY OF 212 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.3	10.0	-1.91	.51	1.08	.32	.85	.09
SEM	.3	.0	.06	.01	.03	.05	.03	.03
P.SD	5.0	.0	.85	.22	.48	.66	.40	.48
S.SD	5.0	.0	.85	.22	.48	.66	.40	.49
MAX.	23.0	10.0	-.08	.97	2.51	2.07	2.07	1.59
MIN.	1.0	10.0	-3.32	.27	.38	-1.49	.20	-1.29

REAL RMSE	.59	TRUE SD	.60	SEPARATION 1.01	Person RELIABILITY	.51
MODEL RMSE	.56	TRUE SD	.64	SEPARATION 1.14	Person RELIABILITY	.57
S.E. OF Person MEAN = .06						

Person RAW SCORE-TO-MEASURE CORRELATION = .93
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .70 SEM = 2.74
STANDARDIZED (50 ITEM) RELIABILITY = .87

SUMMARY OF 10 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	133.1	212.0	.00	.11	.99	.24	.85	-.51
SEM	32.9	.0	.27	.01	.11	.73	.11	.64
P.SD	98.8	.0	.80	.04	.33	2.18	.34	1.93
S.SD	104.2	.0	.84	.04	.35	2.30	.36	2.04
MAX.	370.0	212.0	1.30	.19	1.58	4.32	1.40	3.12
MIN.	25.0	212.0	-1.41	.06	.57	-2.70	.44	-2.50

REAL RMSE	.12	TRUE SD	.79	SEPARATION 6.57	Item RELIABILITY	.98
MODEL RMSE	.12	TRUE SD	.79	SEPARATION 6.80	Item RELIABILITY	.98
S.E. OF Item MEAN = .27						

Fig. 4. Summary statistics of 212 students and 10 items of mathematical literacy questions

Conversely, the item reliability coefficient is very high at 0.98, demonstrating excellent stability in item difficulty estimates. Item reliability indicates the likelihood that the same item difficulty hierarchy would be replicated in another similar sample. Values over 0.90 are generally considered strong evidence of a well-defined item difficulty continuum [26]. This suggests that the mathematical literacy items are consistently understood and effectively spread across the difficulty scale.

The robustness of item reliability is further confirmed by the item separation index of 6.57, which greatly surpasses the minimum recommended value of 3.0. High item separation signifies the presence of multiple distinct levels of item difficulty, allowing

for a confident ranking of items from easiest to hardest. According to Masters, such separation strongly supports adequate construct representation. Additionally, the KR-20 value of 0.70 indicates acceptable internal consistency, complementing the Rasch-based reliability evidence. Overall, these findings suggest that while the instrument shows strong item-level reliability, improvements in test length or targeting are necessary to enhance person-level measurement precision.

4 Conclusion

This study developed and validated a mathematical literacy instrument through expert judgment and empirical testing with students. Expert ratings provided evidence of item quality in terms of ability, process, content, context, and sentence structure, while the Rasch analysis of student responses examined dimensionality, item difficulty, item fit, reliability, and differential item functioning (DIF). The final 10-item instrument showed reasonable evidence of unidimensionality, with 42.7% of the variance explained by the Rasch dimension and a first residual contrast of 1.94 (11.1%). Item reliability (0.98) and separation (6.57) indicated a well-defined hierarchy of item difficulty. Although IN1 and EM3 showed overfit, they were retained because overfit alone was not considered sufficient evidence for item removal. Most items functioned similarly across gender groups, while FO1 showed statistically significant gender-related DIF and requires further examination.

The instrument nevertheless has limitations, particularly its relatively low person reliability (0.51) and person separation (1.01), which indicate limited precision in distinguishing ieindividual students across multiple ability levels. In its current form, the instrument is therefore better suited to group-level assessment and educational research than to high-stakes individual decisions. Future studies should examine the stability of IN1 and EM3, review the gender-related DIF identified for FO1, expand the item pool, and involve larger and more diverse samples to improve measurement precision and provide stronger evidence of the instrument's functioning across different student populations and educational contexts.

Acknowledgment. The authors sincerely thank the experts for their valuable time and thoughtful input during the validation of the mathematical literacy items. We are also grateful to the school principals for their support and assistance during data collection, and to all students who participated in the study. Their cooperation and contributions made this research possible.

Disclosure of Interests. The authors have no competing interests.

References

1. L. A. Steen, R. Turner, and H. Burkhardt, "Developing Mathematical Literacy," vol. 10, pp. 285–294, 2007 2007, doi: 10.1007/978-0-387-29822-1_30.
2. T. R. Hayati and K. Kamid, "Analysis of Mathematical Literacy Processes in High School Students," International Journal of Trends in Mathematics Education Research, vol. 2, no. 3, pp. 116–119, 2019 2019, doi: 10.33122/ijtmer.v2i3.70.
3. S. Ahyan, Tutmudi, and D. Juandi, "Bibliometric analysis of research on mathematical literacy in Indonesia," Journal of Physics: Conference Series, vol. 1869, no. 1, p. 012120, 2021 2021, doi: 10.1088/1742-6596/1869/1/012120.
4. F. O. Haara, O. H. Bolstad, and E. S. Jenssen, "Research on mathematical literacy in schools - Aim, approach and attention," European Journal of Science and Mathematics Education, vol. 5, no. 3, pp. 285–313, 2017 2017, doi: 10.30935/scimath/9512.
5. J. Kilpatrick, "Understanding mathematical literacy: The contribution of research," Educational Studies in Mathematics, vol. 47, no. 1, pp. 101–116, 2001 2001, doi: 10.1023/a:1017973827514.
6. Z. Zulkardi and A. W. Kohar, "Designing PISA-Like Mathematics Tasks In Indonesia: Experiences and Challenges," Journal of Physics: Conference Series, vol. 947, no. 1, p. 012015, 2018/01/01 2018, doi: 10.1088/1742-6596/947/1/012015.
7. S. Ahyan, J. Turmudi, and D, "Mathematical literacy of ninth-grade students in solving PISA mathematics problems," International Journal of Innovation, Creativity and Change, vol. 5, no. 6, pp. 483–95, 2019 2019.
8. D. Khanolainen, Psyridou, M., Silinskas, G., Lerkkanen, M., Niemi, P., Poikkeus, A., & Torppa, M., "Longitudinal Effects of the Home Learning Environment and Parental Difficulties on Reading and Math Development Across Grades 1–9," Frontiers in Psychology, vol. 11, p. 577981, 2020, doi: 10.3389/fpsyg.2020.577981.
9. J.-A. Lefevre, S.-L. Skwarchuk, B. Smith-Chant, L. Fast, D. Kamawar, and J. Bisanz, "Home Numeracy Experiences and Children's Math Performance in the Early School Years," Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, vol. 41, no. 2, pp. 55-66, 2009, doi: 10.1037/a0014532.
10. V. M. Kolar and T. Hodnik, "Mathematical literacy from the perspective of solving contextual problems.," European Journal of Educational Research, vol. 10, no. 1, pp. 467-483, 2021, doi: 10.12973/eu-jer.10.1.467.
11. Ü. Çakıroğlu, M. Güler, M. Dündar, and F. Coşkun, "Virtual Reality in Realistic Mathematics Education to Develop Mathematical Literacy Skills," International Journal of Human–Computer Interaction, vol. 40, no. 17, pp. 4661–4673, 2023 2023, doi: 10.1080/10447318.2023.2219960.
12. S. Sumirattana, A. Makanong, and S. Thipkong, "Using realistic mathematics education and the DAPIC problem-solving process to enhance secondary school students' mathematical literacy," Kasetsart Journal of Social Sciences, vol. 38, no. 3, pp. 307–315, 2017 2017, doi: 10.1016/j.kjss.2016.06.001.
13. T. W. Lee, S. J. Kang, H. J. Lee, and S. I. Hyun, "Testing health literacy skills in older Korean adults," Patient Education and Counseling, vol. 75, no. 3, pp. 302–307, 2009 2009, doi: 10.1016/j.pec.2009.04.002.
14. S. S. Richtering et al., "Examination of an eHealth literacy scale and a health literacy scale in a population with moderate to high cardiovascular risk: Rasch analyses," PLOS ONE, vol. 12, no. 4, p. 0175372, 2017 2017, doi: 10.1371/journal.pone.0175372.

15. O. N. Medvedev, S. D. Pratscher, and A. Bettencourt, "Psychometric Evaluation of the Interpersonal Mindfulness Scale Using Rasch Analysis," Mindfulness, vol. 11, no. 8, pp. 2007–2015, 2020 2020, doi: 10.1007/s12671-020-01415-5.
16. A. Bell and K. Jones, "Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data," Political Science Research and Methods, vol. 3, no. 1, pp. 133–153, 2015 2015, doi: 10.1017/psrm.2014.7.
17. A. Phakiti, P. D. Costa, L. Plonsky, and S. Starfield, The Palgrave Handbook of Applied Linguistics Research Methodology. Palgrave Macmillan, 2018.
18. S. Ahyan, "Survey cross-sectional mengenai resiliensi dan literasi matematis siswa kelas VIII madarasah tsanawiyah di Lombok Timur menggunakan pemodelan Rasch dan analisis multivariate," Universitas Pendidikan Indonesia: Universitas Pendidikan Indonesia, 2023 2023.
19. T. Eckes, "Many-facet Rasch measurement," 2009 2009. [Online]. Available: https://www.researchgate.net/publication/228465956_Many-facet_Rasch_measurement.
20. T. Eckes, Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. Frankfurt am Main: Peter Lang, 2011.
21. S. Ahyan, S. Supiyati, Fahrurroi, and M. N. Hassan, "An application of many-facet Rasch measurement to validate the numeracy test for elementary students," Infinity, vol. 14, no. 4, pp. 861–876, 2025 2025, doi: 10.22460/infinity.v14i4.p861-876.
22. W. J. Boone, J. R. Staver, and M. S. Yale, Rasch analysis in the human science. Springer, 2014.
23. B. Sumintono and W. Widhiarso, Aplikasi model Rasch pada assessment pendidikan. Penerbit Trim Komunikata, 2015.
24. B. Vaughan, J. Mulcahy, and P. McLaughlin, "The dreem, part 2: psychometric properties in an osteopathic student population," BMC Medical Education, vol. 14, no. 1, 2014 2014, doi: 10.1186/1472-6920-14-100.
25. C. Long, H. Wendt, and T. Dunne, "Applying rasch measurement in mathematics education research: steps towards a triangulated investigation into proficiency in the multiplicative conceptual field," Educational Research and Evaluation, vol. 17, no. 5, pp. 387–407, 2011 2011, doi: 10.1080/13803611.2011.632661.
26. T. G. Bond and C. M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences, 3rd ed. Routledge, 2015.
27. J. M. Linacre, "Many-facet Rasch measurement," Journal of Applied Measurement, vol. 5, no. 4, pp. 1–35, 1994 1994.
28. B. D. Wright and G. N. Masters, Rating scale analysis. MESA Press, 1982.
29. G. N. Masters, "A Rasch model for partial credit scoring," Psychometrika, vol. 47, no. 2, pp. 149–174, 1982 1982, doi: 10.1007/BF02296272.
30. OECD, PISA 2018 assessment and analytical framework. OECD Publishing, 2019.
31. J. M. Linacre, "A user's guide to Winsteps® Rasch-model computer programs," Winsteps. com, 2011 2011.
32. J. M. Linacre, "FACETS® Rasch measurement computer program (Version 3.83.6," Winsteps. com, 2023 2023.
33. OECD, PISA 2021 Mathematics Framework. OECD Publishing, 2020.
34. P. W. Holland and H. Wainer, Differential item functioning. Lawrence Erlbaum Associates, 1993.
35. OECD, PISA 2022 Results (Volume I). OECD Publishing, 2023.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

