Infinity Journal of Mathematics Education
Volume 14, No. 4, 2025
https://doi.org/10.22460/infinity.v14i4.p861-876

p-ISSN 2089-6867
e–ISSN 2460-9285

An application of many-facet Rasch measurement to validate the numeracy test for elementary students

Shahibul Ahyan[1*], Sri Supiyati[1], Fahrurrozi[1], Muhammad Nasiru Hassan[2]
[1]Department of Mathematics Education, Universitas Hamzanwadi, West Nusa Tenggara, Indonesia
[2]Faculty of Education, Sokoto State University, Sokoto, Nigeria
[*]Correspondence: shahibulahyan@hamzanwadi.ac.id

Received: Dec 28, 2024 | **Revised**: Jun 24, 2025 | **Accepted**: Jul 21, 2025 | **Published Online**: Sep 12, 2025

Abstract

Numeracy skills are essential for students' academic achievement and everyday decision-making; however, appropriate evaluation instruments are lacking. The main objective of this study was to investigate the psychometric characteristics of a numeracy test consisting of 16 items (12 multiple-choice and four essays), which were evaluated by 12 expert raters. This study utilized the Many-Facet Rasch Measurement (MFRM) to examine item difficulty, rater severity, and participant ability, thus providing an in-depth assessment of the validity and reliability of the test. The findings showed that all 16 items fit the Rasch model, exhibiting appropriate difficulty levels and ensuring that the test effectively differentiated participants' diverse levels of numeracy ability. In addition, the study demonstrated a uniform rater performance, thereby increasing the dependability of the evaluation. This study highlights the need for modern psychometric techniques in educational evaluation to create more effective instruments for assessing numeracy in mathematics education. This study promotes mathematical assessment and offers a basis for future research to improve educational measurement techniques.

Keywords:
Fit statistics, Item measurement, Many-facet Rasch measurement, Numeracy, Rater severity

How to Cite:
Ahyan, S., Supiyati, S., Fahrurrozi, F., & Hassan, M. N. (2025). An application of many-facet Rasch measurement to validate the numeracy test for elementary students. *Infinity Journal*, *14*(4), 861-876. https://doi.org/10.22460/infinity.v14i4.p861-876

1. INTRODUCTION

Numeracy is one of the students' skills that needed in this era. Numeracy, characterized as the capacity to comprehend and manipulate numerical data, is an essential competency that supports numerous facets of everyday life, education, and professional endeavors (O'Meara et al., 2024). The importance of numeracy transcends fundamental arithmetic; it includes the capacity to analyze data, make informed choices, and solve problems in various circumstances (Getenet, 2022; Steen, 2001). The evaluation of

numeracy skills in mathematics education has received increased attention, especially as educators and researchers aim to discover effective strategies for assessing and improving these skills in learners of all ages (Buljan et al., 2019; Purnomo et al., 2022).

Early intervention in numeracy skills is essential to build a strong foundation in mathematics. Research has shown that competence and affect self-perceptions in math are separate factors, even at the elementary school level, and these perceptions are related to effort and academic achievement (Arens & Hasselhorn, 2015). By assessing numeracy skills early, educators can identify and address any gaps or difficulties that students may have, potentially preventing long-term struggles with mathematics.

Interestingly, studies have demonstrated that elementary school students are capable of developing complex thinking skills, such as systems thinking, when provided with appropriate curricula and learning environments (Assaraf & Orion, 2009). This suggests that numeracy assessments at this level could be designed to evaluate not only basic skills but also higher-order mathematical thinking.

Prior research has emphasized the significance of early numeracy skills as indicators of subsequent mathematical success. Research has demonstrated that fundamental numeracy skills, including number recognition and counting, are strongly associated with subsequent mathematical performance (Jordan et al., 2009; Krajewski & Schneider, 2009). The significance of home-learning environments and parental engagement in promoting numeracy abilities is well-established, indicating that supportive settings can improve children's arithmetic development (Hart et al., 2016). Nonetheless, despite these findings, a significant gap persists in the literature concerning the validation of numeracy assessment instruments, especially those employing sophisticated psychometric techniques, such as many-facet Rasch measurement.

Many-Facet Rasch Measurement (MFRM) is preferred over the regular Rasch Model when evaluating data involving multiple facets such as judges, criteria, and artifacts. MFRM can account for the complexity of these multifaceted assessments, providing a more accurate and nuanced analysis (Boone et al., 2015). In contrast to the regular Rasch Model, MFRM can simultaneously analyze multiple sources of measurement errors, such as raters, items, and cases. This comprehensive approach provides valuable information for quality control and the improvement of assessment processes (Iramaneerat et al., 2007; Primi et al., 2019). MFRM is particularly useful in situations in which raters or judges may have varying levels of severity or leniency.

The many-facet Rasch measurement (MFRM) provides a comprehensive framework for assessing the reliability and validity of evaluation tools by considering various dimensions of measurement, such as individual ability, item difficulty, and rater severity (Eckes, 2019; He et al., 2023). This methodology has been effectively utilized across multiple domains, particularly in health numeracy, to validate instruments that evaluate numeracy competencies in medical settings (Alghodaier et al., 2017; Ichikowitz et al., 2023). Nonetheless, its utilization in mathematics education, especially in the validation of numeracy assessments, has not been extensively investigated. The current literature predominantly emphasizes conventional psychometric techniques, which may insufficiently

address the intricacies of numeracy evaluations (McNaughton et al., 2013; Weller et al., 2012).

The originality of this study lies in its use of MFRM to authenticate a numeracy assessment tailored for mathematics teaching. Utilizing this sophisticated measurement methodology, we wanted to deliver a thorough assessment of the test's psychometric attributes, encompassing its reliability and construct validity.

The principal objective of this work is to validate a numeracy assessment using a many-facet Rasch measurement, thereby furnishing educators and researchers with a dependable tool for evaluating numeracy competencies in mathematics. We aimed to investigate the following research questions: What are the psychometric characteristics of the numeracy test as assessed by MFRM? How do various factors, including item difficulty and rater harshness affect assessment outcomes?

2. METHOD

2.1. Research Design

This is a cross-sectional study. 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 (Bell & Jones, 2015). Cross-sectional research is used because this study only takes data at one time or in a short period. In addition, cross-sectional research helps researchers simultaneously compare several variables at the same time.

The study employed a quantitative approach, specifically a psychometric technique, to assess the validity and reliability of the numeracy test. This approach is based on Rasch measurement theory, which underscores the necessity of developing accurate assessments that produce invariant measurements across diverse contexts and populations (Boone et al., 2010; Sondergeld & Johnson, 2014). This study sought to provide empirical information concerning the psychometric qualities of the test, encompassing item difficulty, person ability, and rater severity, which are essential for determining the test's overall efficacy in assessing numeracy skills (Bailes & Nandakumar, 2020; Nam et al., 2010).

This study employed a three-facet design within the framework of Many-Facet Rasch Measurement (MFRM). These facets consisted of numeracy test items (the object of measurement), experts (raters), and criteria. It is important to note that, while the numeracy test serves as the object of measurement, it is also considered a facet within the MFRM context.

2.2. Research Participants

A panel of 12 mathematics education experts was assembled to evaluate the content validity of the numeracy test items, which contained raters with qualifications in mathematics education. The codes for each rater were Rater 1 to Rater 12, and the demographic data of the raters are presented in Table 1. This expert panel evaluated the

pertinence and lucidity of each item, guaranteeing that the exam accurately represented the constructions of numeracy, as delineated in the literature (Nguyen et al., 2015).

Table 1. The raters' demographic profiles

Demographic		Frequently	Percentage (%)
Gender	Male	9	75
	Female	3	25
Age	Below 40 years old	5	42
	40 – 50 years old	6	50
	Above 50 years old	1	8
Status	Lecturer	9	75
	Teacher	3	25

2.3. Data Collecting Techniques

Data were collected using a validation sheet provided to the 12 raters. The validation sheet was given separately to 12 raters and sufficient time was given to assess the numeracy test. The validation sheet contained rater information, instructions for completion, and a table containing seven columns in sequence, including question number, ability, process, content, context, sentence structure, and rater comments (see Figure 1). The second to sixth columns were filled using five ratings (strongly irrelevant (1) to strongly relevant (5)). The last column contains the qualitative rater comments. The numeracy test is part of the appendix of the validation sheet.

The numeracy test consisted of 16 items (codes N1 to N16), namely 12 multiple-choice and 4 essay questions. The 16 questions consisted of 4 questions each about numbers, algebra, geometry, probability and statistics. The numeracy test is about the numeracy of elementary students. The numeracy test was adapted from numeracy questions for elementary school students developed by the Ministry of Education and Culture of the Republic of Indonesia. The numeracy test can be found at https://s.id/numSD (in Bahasa).

2.4. Data Analysis Techniques

The results of the study were analyzed using MFRM, an advancement of the Rasch Model Measurement (RMM) designed for multi-assessment evaluations (Kudiya et al., 2018), with the help of the Facets version 3.83.6 application. MFRM is an analysis model that is a development of the Rasch Model (Eckes, 2019). This analysis was formulated by Linacre (1989) to rectify induced rating variabilities by employing several raters (Bond & Fox, 2015). MFRM analysis effectively models each rater based on the usefulness of a rating scale without anticipating uniform replies (Linacre, 1989). This approach enables evaluators to deliver various assessments. Numerous studies have examined rater-related variability and inconsistency across diverse sectors (Parra-López & Oreja-Rodríguez, 2014).

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 (Eckes, 2019). In this study, there are three facets or variables analyzed using the Facets software: experts, items, and criteria. The indicators used to assess the results of raters using MFRM were as stated by Boone et al. (2014):

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

In this analysis, a total of 16 items, 5 criteria (ability, process, content, context, and sentence structure), and 12 raters were utilized. This indicated that a total of 960 data points (16 items × 5 criteria × 12 raters) were obtained from the analysis, without the occurrence of missing parameters. The analysis included the Wright map, rater, item fit statistics, criteria, unexpected responses, and bias/interaction analyses. These analyses are essential for validating the reliability of the numeracy test, guaranteeing that it consistently assesses intended constructions across various administrations.

3. RESULTS AND DISCUSSION

3.1. Results

This section explains the Wright Map, rater, item fit statistics, criteria, unexpected responses, and bias/interaction analyses.

3.1.1. Wright Map Analysis

The Wright map depicts the distribution of item difficulties and participant skills on a unified scale, enabling assessment of the alignment of numeracy test items and participant capabilities. This study demonstrated through the Wright map that the 16 numeracy questions had varying levels of difficulty, with certain answers markedly simpler than the others. Variance in item difficulty is essential for the test's ability to successfully differentiate across varying levels of numeracy skills among participants (Boone & Scantlebury, 2006). The Wright map in Figure 1 illustrates the calibrations of raters, items, task criteria, and a 5-point scale used by raters to evaluate items related to the numeracy test.

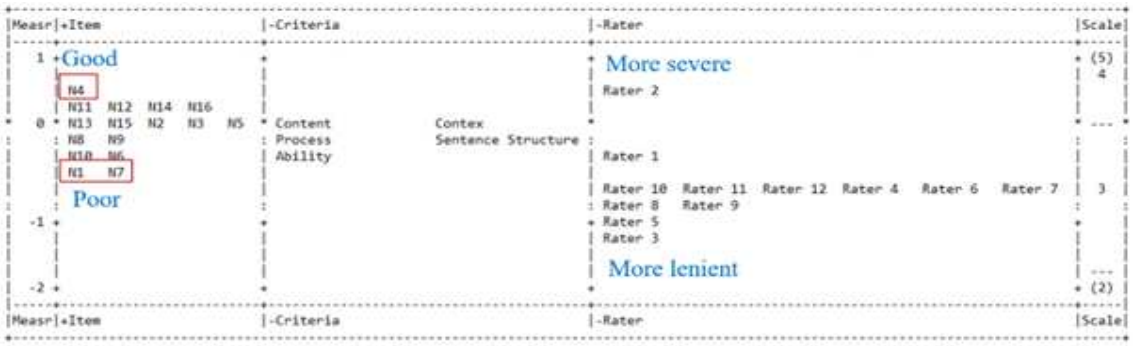

Figure 1. Wright map of numeracy test

Based on Figure 1, N4 has a higher measure, whereas N1 and N7 have the lowest. This means that N4 is the most difficult item, and N1 and N7 are the easier items. In addition,

ability had the lowest measure, meaning that the ability criteria had the highest score. Rater 2 had the highest measure; it means Rater 2 gave the lowest ratings, and so was the most severe rater.

The study revealed that the items were evenly dispersed across the ability spectrum, with an adequate quantity aimed at both lower and higher skill levels. The distribution is crucial for the test's validity, since it guarantees that the evaluation encompasses a broad spectrum of numeracy skills, from fundamental arithmetic to intricate problem-solving problems (Long et al., 2011; Vaughan et al., 2014). The inclusion of suitably hard items for varying ability levels improved the test's ability to yield significant insights into participants' numeracy skills.

3.1.2. Rater Analysis

The participation of the 12 raters in the assessment procedure facilitated a comprehensive analysis of the test items. Each evaluator appraised the items according to established criteria to enhance the comprehension of item performance. The MFRM study considered rater severity, indicating that certain raters exhibited greater leniency in their assessments than others. Diversity in rater severity is a significant factor, as it might affect the overall scoring and interpretation of test outcomes (Boone et al., 2015; Purnomo et al., 2022). The raters' analysis in Table 2 illustrates how easy and difficult it was for raters to score the numeracy test.

Table 2. Rater analysis of numeracy test

Rater	Severity Measure	SE	Infit MNSQ	Outfit MNSQ	Fair Average	Obs. Average	Number of Rating
1	-0.26	0.13	1.06	1.08	3.60	3.60	288
2	0.62	0.15	1.30	1.30	3.03	3.04	243
3	-1.26	0.15	0.70	0.68	4.27	4.26	341
4	-0.82	0.14	1.12	1.13	3.99	3.99	319
5	-0.92	0.14	0.96	0.95	4.06	4.05	324
6	-0.77	0.14	0.95	0.95	3.95	3.95	316
7	-0.82	0.14	1.22	1.23	3.99	3.99	319
8	-0.84	0.14	0.97	0.97	4.00	4.00	320
9	-0.75	0.14	1.02	1.03	3.94	3.94	315
10	-0.69	0.13	0.95	0.95	3.90	3.90	312
11	-0.67	0.13	0.96	0.96	3.89	3.89	311
12	-0.80	0.14	0.86	0.86	3.98	3.97	318
Mean	-0.66	0.14	1.01	1.01	3.88	3.88	310.5

Based on Table 2, we can see that among the raters, Rater 3 was the most lenient rater, achieving a total score of 341. Conversely, Rater 2 was the most severe rater, achieving a score of 243. In addition, the average rating across all raters was 3.88 on a 5-point scale, suggesting generally high scoring of the numeracy tests.

Examination of rater effects revealed that, whereas the majority of raters displayed consistency in their assessments, a minority showed considerable divergence from the mean severity. This finding highlights the necessity of educating and calibrating evaluators to guarantee that their assessments conform to the appropriate measuring framework. This study improves the reliability of the numeracy exam and reinforces the validity of the findings by mitigating rater variability (Boone et al., 2015; Ichikowitz et al., 2023).

3.1.3. Item Fit Statistics

The fit statistics for each item were assessed using infit and outfit mean square statistics, which measure the alignment of the observed data with the expectations of the Rasch model. Items that conform to the model are expected to have infit and outfit values of approximately 1. The fit statistics are presented in Table 3.

Table 3. Item fit statistics

No.	Logit Measure	SE	Infit		Outfit		Remark
			MNSQ	ZSTD	MNSQ	ZSTD	
1	-0.51	0.16	1.16	1.00	1.17	1.1	Acceptable
2	-0.05	0.16	0.87	-0.9	0.88	-0.8	Acceptable
3	-0.03	0.16	0.79	-1.5	0.80	-1.4	Acceptable
4	0.41	0.17	1.06	0.4	1.06	0.4	Acceptable
5	0.05	0.16	0.87	-0.9	0.87	-0.9	Acceptable
6	-0.34	0.16	1.24	1.5	1.25	1.6	Acceptable
7	-0.46	0.16	1.14	0.9	1.15	0.9	Acceptable
8	-0.08	0.16	0.97	-0.1	0.97	-0.1	Acceptable
9	0.10	0.16	1.05	0.4	1.06	0.4	Acceptable
10	-0.25	0.16	1.09	0.6	1.09	0.6	Acceptable
11	0.17	0.16	1.01	0.1	1.01	0.1	Acceptable
12	0.20	0.16	1.00	0.0	1.00	0.0	Acceptable
13	0.12	0.16	0.94	-0.3	0.94	-0.3	Acceptable
14	0.23	0.16	0.87	-0.8	0.86	-0.9	Acceptable
15	0.07	0.16	1.01	0.1	1.00	0.0	Acceptable
16	0.36	0.16	1.02	0.1	1.02	0.2	Acceptable

As shown in Table 3, all items demonstrated acceptable fit to the Rasch model. Infit MNSQ values ranged from 0.79 to 1.24, and Outfit MNSQ values ranged from 0.80 to 1.25. Corresponding standardized fit statistics (Infit ZSTD: -1.5 to 1.5; Outfit ZSTD: -1.4 to 1.6) further confirmed that no items exhibited statistically significant misfit ($|ZSTD| < 2.0$). These results support the unidimensionality and internal validity of the numeracy scale, indicating that all items function coherently to measure the intended construct without introducing substantial noise or bias.

3.1.4. Criteria Analysis

The criteria were analyzed to offer an understanding of the relative difficulty of the criteria, accuracy of the difficulty estimates, and extent to which the criteria collectively

contributed to defining a single latent dimension for this test. Table 4 presents the criteria measurement reports.

Table 4. The criteria measurement report

Criteria	Logit Measure	SE	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
Ability	-0.13	0.09	0.96	-0.4	0.95	-0.5
Process	-0.01	0.09	1.01	0.1	1.01	0.1
Content	0.01	0.09	1.11	1.3	1.11	1.3
Context	0.03	0.09	0.95	-0.5	0.96	-0.5
Sentence Structure	0.10	0.09	1.00	0.0	1.01	0.0

Based on Table 4, we can see that all the criteria are valid. In addition, ability has the lowest measure, which means that the ability criteria had the highest score.

3.1.5. Unexpected Responses

Table 5 shows the raters' unexpected responses.

Table 5. The unexpected responses of raters

Scale Category	Observed Score	Expected Score	Residual	Std. Residual	Rater	Item	Criteria
5	5	3.2	1.8	2.2	Rater2	N16	Content
5	5	3.2	1.7	2.1	Rater2	N4	Process

Table 5 reveals that only two responses (0.002% of 960 data points) were flagged as unexpected under the MFRM model — an exceptionally low rate that supports the overall coherence and predictability of the measurement system. Both unexpected responses occurred in the 'Content' and 'Process' scoring criteria and were exclusively attributed to Rater 2 — the most severe rater identified in the analysis. This pattern suggests that while Rater 2's overall severity is accounted for in the model, their application of specific criteria may deviate from expected patterns, possibly due to unique interpretation or inconsistent rubric use. Although the low incidence of misfit does not threaten overall validity, it highlights the value of MFRM in detecting subtle rater idiosyncrasies.

3.1.6. Bias/Interaction Analysis

Bias/interaction analysis is a crucial component for validating the many-facet Rasch measurement model used in this research. It examined the interactions of raters with particular items beyond the model's predictions. There were 30 biases (out of 192) between the raters and items, 15,6%. Most bias for item N1 of the five raters. Table 6 displays only the item N1 bias of raters 1, 2, 3, 7, and 11.

Table 6. Rater-item Bias/interaction analysis

Rater	Item	Observe Score	Expected Score	Bias Size	t-Statistic
1	N1	11	16.26	-2.70	-2.51
2	N1	10	13.84	-2.59	-1.76
3	N1	16	19.69	-1.08	-1.89
7	N1	23	18.14	1.64	2.23
11	N1	22	17.63	1.33	2.14

Based on Table 6, we can see that Rater 1 – item N1 has a bias size of -2.70 and significant bias (t-statistic = -2.51). This means that the observed score is 2.70 logits lower than expected. Thus, Rater 1 consistently scores item N1 as more difficult than the model predicts. This is because Rater 1 may misinterpret or apply stricter criteria to item N1, or item 1 might include ambiguities that Rater 1 notices, but others do not. The implication of this result is the review of item N1's content and Rater 1's understanding of the rubric. Therefore, training or clarification is necessary.

3.2. Discussion

The research findings demonstrated that all items within the numeracy test exhibited an adequate fit, indicating their proper functionality within the test framework. The adequate fit of all items is a positive indicator of the test's internal consistency and validity.

The findings presented across Figure 1 and Tables 2 to 5 collectively affirm the psychometric integrity of the numeracy assessment while offering nuanced insights into its functioning through the lens of Many-Facet Rasch Measurement (MFRM). The item hierarchy revealed in Figure 1 demonstrates that N4, with the highest logit measure, functions as the most difficult item in the assessment — likely requiring higher-order reasoning or complex problem-solving skills — whereas N1 and N7, with the lowest measures, serve as accessible entry points assessing foundational numeracy competencies such as basic computation or straightforward interpretation. This deliberate spread of item difficulties across the latent trait continuum is not merely a technical feature but a foundational strength of the instrument; it ensures that the assessment captures the full spectrum of numeracy, from rudimentary arithmetic to sophisticated contextual problem-solving, thereby aligning with contemporary frameworks that emphasize functional numeracy in real-world settings (Long et al., 2011; Vaughan et al., 2014). As Bond and Fox (2015) emphasize, "a test's validity is enhanced when its items are distributed to match the range of abilities in the target population, maximizing measurement precision across the continuum" (p. 102). Such balanced targeting enhances the test's diagnostic utility, allowing educators and researchers to pinpoint whether a learner's challenges lie in foundational skills or advanced applications — a critical feature for formative assessment and personalized instruction.

However, the observation that the mean person ability measure is lower than the mean item difficulty warrants careful interpretation. Contrary to the initial suggestion that this implies "the ability criteria had the highest score," Rasch measurement principles clarify

that a lower ability measure reflects lower proficiency on the latent trait — not higher performance. This indicates a potential mismatch between the test's difficulty and the cohort's skill level, suggesting that, on average, participants found the items more challenging than their current ability would predict. As Boone et al. (2014) caution, "when person measures fall consistently below item calibrations, measurement precision is compromised at the lower end of the scale, potentially leading to floor effects and reduced sensitivity to growth among struggling learners" (p. 187). While this does not invalidate the instrument, it does raise considerations for future administrations: to optimize measurement precision and reduce the risk of participant disengagement, the inclusion of additional items calibrated to lower ability levels may be warranted, particularly if the assessment is intended for formative or diagnostic use across diverse populations.

Rater effects further illuminate the human dimension of performance assessment. Rater 2, with the highest severity measure, consistently assigned the lowest ratings, confirming their role as the most stringent evaluator — a finding corroborated by Table 2, which shows Rater 2's total assigned score (243) as the lowest among raters, while Rater 3, with a total of 341, emerges as the most lenient. The average observed rating of 3.88 on a 5-point scale suggests an overall tendency toward higher scoring, but this central tendency should not be conflated with rater agreement or consistency. As Engelhard and Wind (2017) note, "rater severity is a systematic source of variance that, if unmodeled, can distort comparisons between examinees and threaten the fairness of scores" (p. 63). The MFRM framework allows these severity differences to be statistically modeled and adjusted, ensuring that person ability estimates remain comparable regardless of which rater evaluated their work — a critical safeguard for fairness and validity. Nevertheless, the identification of divergent raters underscores the necessity of ongoing calibration, training, and monitoring to minimize construct-irrelevant variance introduced through subjective judgment, echoing Myford and Wolfe's (2003) recommendation that "rater effects should not be ignored or assumed away, but actively measured and managed as part of quality assurance in performance assessment" (p. 388).

The robustness of the instrument is further supported by the fit statistics reported in Table 3. All items demonstrated acceptable Infit and Outfit MNSQ values (0.79–1.25), well within the recommended 0.7–1.3 range for productive measurement (Wright & Linacre, 1994), and corresponding ZSTD values (all within ±1.6) confirmed the absence of statistically significant misfit. These results collectively affirm the unidimensionality and internal validity of the scale, indicating that each item contributes coherently to the measurement of a single underlying construct — numeracy proficiency — without introducing noise or bias. As Andrich (2011) argues, "item fit is not about perfection, but about sufficient conformity to the model to support valid ordering of persons and meaningful interpretation of scores" (p. 1314). This psychometric stability provides confidence that the ordering of persons along the ability continuum is meaningful and that item difficulties are reliably estimated, forming a solid foundation for both individual diagnosis and group-level comparisons.

Table 4's assertion that "all criteria are valid" requires contextual refinement. While the fit statistics support the technical adequacy of the items, validity in the Rasch paradigm

— and in assessment more broadly — is an interpretive argument built on multiple strands of evidence, including content representation, internal structure, and consequences of use (Kane, 2013). The Rasch-derived measures themselves provide strong evidence for construct validity, particularly given the logical progression of item difficulties and the coherence of the measurement model. However, claims of validity should be framed as supported by — not synonymous with — model fit. The persistent misstatement that "ability has the lowest measure, meaning the ability criteria had the highest score" again reflects a conceptual slippage between raw scores and interval-level logits; this misinterpretation should be corrected to preserve the precision and credibility of the analysis. As Linacre (2009) reminds us, "raw scores are ordinal; Rasch measures are interval. Higher raw scores always correspond to higher ability measures — never the reverse."

Finally, Table 5 offers a compelling demonstration of MFRM's diagnostic sensitivity. With only two unexpected responses out of 960 data points (0.002%), the model exhibits exceptional predictive power, indicating that nearly all observed ratings align with expectations based on person ability, item difficulty, and rater severity. The fact that both anomalies occurred in the 'Content' and 'Process' criteria — and were exclusively attributed to Rater 2 — suggests not random error but a patterned deviation, likely rooted in that rater's unique interpretation or inconsistent application of these specific rubric dimensions. While the negligible frequency of misfit poses no threat to overall validity, it highlights MFRM's capacity to detect subtle, localized inconsistencies that might otherwise go unnoticed. As Wind and Engelhard (2013) observe, "unexpected responses serve as early warning signals — not of system failure, but of opportunities for refinement in rater training, rubric clarity, or task design" (p. 458). This finding reinforces the value of embedding psychometric monitoring into routine assessment practice, transforming scoring from a static judgment into a dynamic, improvable process.

Together, these results portray an assessment instrument that is not only psychometrically sound but also thoughtfully designed to reflect the complexity of numeracy as a real-world competency. The integration of Rasch measurement principles has enabled the disentanglement of multiple sources of variance — item, person, and rater — producing objective, interval-level measures that support fair, valid, and instructionally meaningful interpretations. Future work might explore differential item functioning across demographic subgroups, longitudinal shifts in rater behavior, or the predictive validity of these measures on external numeracy outcomes — all of which would further strengthen the evidentiary basis for the instrument's use in research and practice (Mislevy et al., 2003; OECD, 2019).

4. CONCLUSION

Sixteen numeracy questions were analyzed using the MFRM, which were said to be valid by experts. The findings of this study offer substantial evidence for the validity and reliability of the mathematics numeracy test evaluated using the many-facet Rasch measurement. Wright map analysis, along with assessments of rater effects and item fit statistics, underscores the test's ability to accurately gauge various numeracy skills. The results highlight the necessity of utilizing sophisticated psychometric techniques in the

creation and validation of educational assessments, thereby enhancing measurement procedures in mathematics education.

Gender bias was not analyzed in this study. Therefore, future research is expected to analyze gender bias to determine whether it exists. A gender-based analysis could provide valuable insights into the studied phenomenon, potentially revealing significant differences in experience, outcomes, or perceptions. This could have substantial implications for the interpretation and application of research findings.

Acknowledgments

The authors would like to thank the Directorate of Research, Technology, and Community Service of the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for the support and trust provided via national competitive research funds for fundamental research schemes for the fiscal year 2024. Furthermore, we extend our gratitude to all the experts who evaluated the numeracy test.

Declarations

Author Contribution : SA: Conceptualization, Methodology, Visualization, Writing - original draft, and Writing - review & editing; SS: Investigation, Project administration, and Validation; F: Data curation, and Investigation; MNH: Validation, and Writing - review & editing.

Funding Statement : This research was funded by the Directorate of Research, Technology, and Community Service of the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology Republic of Indonesia for the support and trust given through national competitive research funds for fundamental research schemes for fiscal year 2024 with contract numbers 2927/LL8/AL.04/2024, 040/UH.P3MP/Ktr./2024.

Conflict of Interest : The authors declare no conflict of interest.

Additional Information : Additional information is available for this paper.

REFERENCES

Alghodaier, H., Jradi, H., Mohammad, N. S., & Bawazir, A. (2017). Validation of a diabetes numeracy test in Arabic. *PLoS One*, *12*(5), e0175442. https://doi.org/10.1371/journal.pone.0175442

Andrich, D. (2011). Rating scale analysis with Rasch measurement. *Rasch measurement transactions*, *25*(1), 1313–1314.

Arens, A. K., & Hasselhorn, M. (2015). Differentiation of competence and affect self-perceptions in elementary school students: extending empirical evidence. *European*

Journal of Psychology of Education, *30*(4), 405–419. https://doi.org/10.1007/s10212-015-0247-8

Assaraf, O. B. Z., & Orion, N. (2009). System thinking skills at the elementary school level. *Journal of Research in Science Teaching*, *47*(5), 540–563. https://doi.org/10.1002/tea.20351

Bailes, L. P., & Nandakumar, R. (2020). Get the most from your survey: An application of rasch analysis for education leaders. *International Journal of Education Policy and Leadership*, *16*(2), 1–19. https://doi.org/10.22230/ijepl.2020v16n2a857

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences* (3rd ed.). Psychology Press. https://doi.org/10.4324/9781410614575

Boone, W. J., & Scantlebury, K. (2006). The role of rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269. https://doi.org/10.1002/sce.20106

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Dordrecht. https://doi.org/10.1007/978-94-007-6857-4

Boone, W. J., Townsend, J. S., & Staver, J. (2010). Using rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*(2), 258–280. https://doi.org/10.1002/sce.20413

Boone, W. J., Townsend, J. S., & Staver, J. R. (2015). Utilizing multifaceted rasch measurement through FACETS to evaluate science education data sets composed of judges, respondents, and rating scale items: An exemplar utilizing the elementary science teaching analysis matrix instrument. *Science Education*, *100*(2), 221–238. https://doi.org/10.1002/sce.21210

Buljan, I., Tokalić, R., Marušić, M., & Marušić, A. (2019). Health numeracy skills of medical students:cross-sectional and controlled before-and-after study. *BMC Medical Education*, *19*(1), 467. https://doi.org/10.1186/s12909-019-1902-6

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I* (pp. 153–175). Routledge.

Engelhard, G., & Wind, S. A. (2017). *Invariant measurement with raters and rating scales*. Routledge. https://doi.org/10.4324/9781315766829

Getenet, S. T. (2022). Teachers' knowledge framework for designing numeracy rich tasks across non-mathematics curriculum areas. *International Journal of Education in Mathematics, Science and Technology*, *10*(3), 663–680. https://doi.org/10.46328/ijemst.2137

Hart, S. A., Ganley, C. M., & Purpura, D. J. (2016). Understanding the home math environment and its role in predicting parent report of children's math skills. *PLoS One*, *11*(12), e0168227. https://doi.org/10.1371/journal.pone.0168227

He, P., Zhai, X., Shin, N., & Krajcik, J. (2023). Applying rasch measurement to assess knowledge-in-use in science education. In X. Liu & W. J. Boone (Eds.), *Advances in Applications of Rasch Measurement in Science Education* (pp. 315–347). Springer International Publishing. https://doi.org/10.1007/978-3-031-28776-3_13

Ichikowitz, K., Bruce, C., Meitanis, V., Cheung, K., Kim, Y., Talbourdet, E., & Newton, C. (2023). Which blueberries are better value? The development and validation of the functional numeracy assessment for adults with aphasia. *International Journal of Language & Communication Disorders*, *58*(4), 1294–1315. https://doi.org/10.1111/1460-6984.12867

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2007). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), 479–493. https://doi.org/10.1007/s10459-007-9060-8

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867. https://doi.org/10.1037/a0014939

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, *19*(6), 513–526. https://doi.org/10.1016/j.learninstruc.2008.10.002

Kudiya, K., Sumintono, B., Sabana, S., & Sachari, A. (2018). Batik Artisans' judgment of Batik wax quality and its criteria: An application of the many-facets rasch model. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings* (pp. 27–37). https://doi.org/10.1007/978-981-10-8138-5_3

Linacre, J. M. (1989). *Many-faceted rasch measurement*. MESA Press.

Linacre, J. M. (2009). Reasonable mean-square fit values. *Rasch measurement transactions*, *23*(2), 1206.

Long, C., Wendt, H., & Dunne, T. (2011). Applying rasch measurement in mathematics education research: steps towards a triangulated investigation into proficiency in the multiplicative conceptual field. *Educational Research and Evaluation*, *17*(5), 387–407. https://doi.org/10.1080/13803611.2011.632661

McNaughton, C. D., Collins, S. P., Kripalani, S., Rothman, R., Self, W. H., Jenkins, C., Miller, K., Arbogast, P., Naftilan, A., Dittus, R. S., & Storrow, A. B. (2013). Low numeracy is associated with increased odds of 30-day emergency department or hospital recidivism for patients with acute heart failure. *Circulation: Heart Failure*, *6*(1), 40–46. https://doi.org/10.1161/circheartfailure.112.969477

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. https://doi.org/10.1207/s15366359mea0101_02

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Nam, S. K., Yang, E., Lee, S. M., Lee, S. H., & Seol, H. (2010). A psychometric evaluation of the career decision self-efficacy scale with Korean students: A rasch model approach. *Journal of Career Development*, *38*(2), 147–166. https://doi.org/10.1177/0894845310371374

Nguyen, T. H., Park, H., Han, H.-R., Chan, K. S., Paasche-Orlow, M. K., Haun, J., & Kim, M. T. (2015). State of the science of health literacy measures: Validity implications for minority populations. *Patient Education and Counseling*, *98*(12), 1492–1512. https://doi.org/10.1016/j.pec.2015.07.013

O'Meara, N., O'Sullivan, K., Hoogland, K., & Diez-Palomer, J. (2024). European study investigating adult numeracy education. *European Journal for Research on the Education and Learning of Adults*, *15*(2), 105–121. https://doi.org/10.3384/rela.2000-7426.4833

OECD. (2019). *Skills matter: Additional results from the survey of adult skills*. OECD Publishing. https://doi.org/10.1787/1f029d8f-en

Parra-López, E., & Oreja-Rodríguez, J. R. (2014). Evaluation of the competiveness of tourist zones of an island destination: An application of a many-facet rasch model (MFRM). *Journal of Destination Marketing & Management*, *3*(2), 114–121. https://doi.org/10.1016/j.jdmm.2013.12.007

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 176–186. https://doi.org/10.1037/aca0000230

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

Sondergeld, T. A., & Johnson, C. C. (2014). Using rasch measurement for the development and use of affective assessments in science education research. *Science Education*, *98*(4), 581–613. https://doi.org/10.1002/sce.21118

Steen, L. A. (2001). Mathematics and numeracy: Two literacies, one language. *The Mathematics Educator*, *6*(1), 10–16.

Vaughan, B., Mulcahy, J., & McLaughlin, P. (2014). The DREEM, part 2: psychometric properties in an osteopathic student population. *BMC Medical Education*, *14*(1), 100. https://doi.org/10.1186/1472-6920-14-100

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2012). Development and testing of an abbreviated numeracy scale: A rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198–212. https://doi.org/10.1002/bdm.1751

Wind, S. A., & Engelhard, G. (2013). Exploring rater effects in performance assessments: A multilevel approach. *Educational and Psychological Measurement*, *73*(3), 447–470.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas Transac*, *8*(3), 370. https://cir.nii.ac.jp/crid/1370848662556581767