Elementary Education Journal, 5(1), 2025, 40-50
Available at: https://journal.nurscienceinstitute.id/index.php/eej
EISSN: 2809-4689

Validation of an Ethnomathematics-Based Numeracy Literacy Assessment in the Javanese Cultural Context for Elementary School Students

Nur Romdlon Maslahul Adi[1*], Fina Wardani[2], Isna laili Qurroti A'yun[3]
[1,2]UIN Sunan Ampel Surabaya, Surakarta, Indonesia
[3]Universitas Diponegoro, Semarang, Indonesia
*Email: nur.romdlon.maslahul.adi@uinsa.ac.id

Submitted: 2025-07-26
Accepted: 2025-12-02
Published: 2025-12-27

DOI: 10.53088/eej.v5i1.3594

Keywords:

Numeracy literacy

Ethnomathematics

Instrument validation

Islamic elementary schoool

Madrasah Ibtidaiyah

Abstract

Background: *Poor numeracy performance among Indonesian students highlights the need for assessment tools with strong psychometric properties that are also culturally relevant. This study examined the validity and reliability of a numeracy assessment instrument based on Javanese ethnomathematics for fifth-grade students at an Islamic elementary school (Madrasah Ibtidaiyah).*
Method: *The article reports the quantitative validation and psychometric testing phases of a broader Research and Development (R&D) study based on the Borg and Gall model. Thirty test items were developed in multiple-choice, complex multiple-choice, true-false, and short-answer formats, covering content domains, cognitive processes, and Javanese cultural contexts. Content validity was evaluated using Aiken's V and the Intraclass Correlation Coefficient (ICC), while Rasch modeling was applied to examine the instrument's overall psychometric characteristics.*
Result: *The findings showed that all stimuli met the content validity criterion (Aiken's V ≥ 0.80), and inter-rater reliability was excellent (ICC = 0.909). Rasch analysis confirmed unidimensionality (raw variance explained = 46.0%), all items demonstrated positive polarity, and only one item failed to meet the item-fit criterion. Item reliability in the limited trial was very high (0.96). These findings indicate that the instrument possesses adequate psychometric quality to measure primary school students' numeracy reliably within a Javanese cultural context.*
Implication: *The validated instrument provides a culturally responsive tool for assessing primary school students' numeracy within a Javanese cultural context.*

INTRODUCTION

Numeracy literacy is currently recognized as an essential foundation in the 21st century. Through this skill, students are enabled to think critically, solve problems, and make appropriate decisions in their daily lives. Furthermore, numeracy mastery is used as a key indicator for achieving the fourth Sustainable Development Goals (SDGs) regarding quality education. Despite its undeniable importance, the low numeracy achievement among Indonesian students remains a serious issue that demands immediate attention.

Within the context of basic education, numeracy is not simply restricted to understanding basic arithmetic activities. Logical reasoning, the application of mathematical concepts to real-world situations, and the ability to communicate findings are also highly required. Therefore, strengthening numeracy literacy must be prioritized to improve overall educational quality, particularly in Islamic Elementary Schools (Madrasah Ibtidaiyah). This strategic measure is considered necessary because such skills will assist as the fundamental basis for learning mathematics at higher educational levels.

However, the contemporary state of numeracy achievement among Indonesian students remains a significant concern. The 2022 Programme for International Student Assessment (PISA) report showed the average mathematical literacy score for these students reached only 366. This figure falls substantially below the OECD average of 472, positioning Indonesia at 70th out of 81 participating nations (OECD, 2023). While the country experienced a slight improvement in its overall ranking compared to the 2018 cycle, the actual performance score dropped by 13 points.

Ethnomathematics has emerged as a promising approach to bridge the existing numeracy gap. This framework refers to the mathematical practices that naturally evolve within specific cultural groups (D'Ambrosio, 1985). By leveraging ethnomathematics, school mathematics can be directly connected to students' lived cultural experiences, making learning inherently more meaningful (Rosa & Orey, 2023). Pedagogically, this cultural integration humanizes mathematics, transforming it from an abstract discipline into a relatable concept. Nevertheless, its practical application faces ongoing challenges, particularly concerning the development of genuinely contextual assessment tools (Batiibwe, 2025).

The positive impacts of ethnomathematics on learning outcomes have been widely documented. Prahmana and D'Ambrosio (2020) reported that recognizing the interplay between mathematics and culture significantly reduces math anxiety while enhancing student engagement. Furthermore, improvements in conceptual understanding and primary-school numeracy skills are consistently supported by culturally responsive instruction (Payadnya et al., 2024; Solihin et al., 2025). Despite these promising findings, previous scholarship remains largely concentrated on instructional models and learning materials. Empirical research focusing on ethnomathematics-based assessment tools that possess rigorous psychometric properties remains critically limited (Hidayat Ismail et al., 2023).

Efforts to develop similar assessment instruments have actually been carried out extensively, but their exploration is still dominated by cultural contexts. For example, previous studies have highlighted agricultural ethnomathematics (Safrida et al., 2024), the culture of Musi Rawas (Utari et al., 2024), local wisdom of Banten (Wirawan et al., 2024), and traditions in Jambi and Kuningan (Dewi & Defitriani, 2024; Prayitno et al., 2023). Although the contributions of these various studies are commendable, the testing approaches they use tend to vary and are partial.

On the other hand, the exploration of ethnomathematics that specifically delves into Javanese cultural values, especially those tailored to the characteristics of Madrasah Ibtidaiyah (MI) students, is still very limited. However, the population of students in the Javanese cultural environment is very large, and they need numeracy measurement tools that are relevant to their everyday sociocultural realities. Furthermore, as far as the literature review goes, no study has yet faced up to integrating all aspects of testing comprehensively. Crucial aspects such as content validity, expert agreement consistency, construct validity, and item characteristic analysis have never been integrated into a single psychometric evaluation framework. This research gap must be bridged immediately. The development of numeracy literacy instruments should not stop at the claim of being culturally contextual yet requires empirical evidence from solid validity and reliability testing.

To address this gap, the present study developed and evaluated a numeracy literacy instrument based on Javanese ethnomathematics for fifth-grade Madrasah Ibtidaiyah students. The instrument was validated using multiple psychometric approaches, including expert judgment quantified by Aiken's V and the Intraclass Correlation Coefficient (ICC), construct validity analysis, and Rasch modeling to examine item characteristics. Combining these complementary methods provides more comprehensive evidence of the instrument's validity and reliability.

This study offers both practical and theoretical contributions. Practically, it provides educators with a valid and reliable tool for culturally responsive numeracy assessment. Theoretically, it extends the psychometric literature on ethnomathematics-based assessment while supporting the integration of local wisdom into mathematics education. These contributions align with the principles of the Merdeka Curriculum, which emphasizes contextual learning (Prahmana, 2022), and with evidence that culturally responsive mathematics instruction can enhance students' mathematical competence (Mahmudah & Arif, 2022). Overall, through a systematic process of instrument development, validation, and field testing, this study provides empirical support for implementing culturally responsive numeracy assessment in primary education.

METHOD

This research constitutes quantitative psychometric testing of a macro Research and Development (R&D) study that adopts the Borg and Gall model (Gall et al., 1983). A limited-scale trial was conducted on 30 fifth-grade students of Madrasah Ibtidaiyah in Surabaya City, selected through purposive sampling.

The developed instrument contains 30 items with various formats, including multiple choice, complex multiple choice, true-false, and short answer. The design of this measurement tool is precisely mapped based on three main domains: mathematical content (numbers, algebra, geometry, and data and uncertainty), cognitive processes (understanding, application, and reasoning), and the Javanese ethnomathematics context. The integration of local culture is realized through the representation of batik motifs, temple architecture, traditional dances, and distinctive cuisine.

To achieve high quality in the instrument, a layered validation procedure is applied. Content validation is evaluated by a panel of experts consisting of measurement specialists, mathematicians, and elementary education practitioners. The aspects reviewed include the accuracy of the goal formulation, the relevance of the model, the editorial feasibility, and the guidelines for score interpretation. Expert assessments were then quantified using Aiken's V index, where an item is considered valid if it exceeds the required minimum threshold (Aiken, 1985). Next, the level of inter-rater agreement (inter-rater reliability) was measured using the Intraclass Correlation Coefficient (ICC). Because all validators assessed the same set of items, the analysis used a two-way random effects approach with absolute agreement (Mehta et al., 2018). The reliability classification is as follows: < 0.40 indicates poor reliability, 0.40–0.59 moderate reliability, 0.60–0.74 good reliability, and 0.75–1.00 excellent reliability.

In the final stage, the psychometric characteristics of the items are further analyzed using Rasch modeling (Item Response Theory) using Winsteps software. The evaluation parameters include unidimensionality testing, reliability, construct analysis, and item fit. These fit criteria are based on the INFIT/OUTFIT Mean Square and Point Measure Correlation values. Overall, the triangulation of methods using Aiken's V, ICC, Confirmatory Factor Analysis (CFA), and the Rasch model is implemented to produce comprehensive and layered evidence of validity and reliability.

RESULT

The assessment instrument in this study is designed by combining two main pillars. The first pillar is the Minimum Competency Assessment (AKM) standards from the Ministry of Education and Culture. Furthermore, this national framework is strictly aligned with international measurement standards, namely PISA and TIMSS. This dual alignment becomes the main advantage of the developed instrument. Through it, researchers ensure that the localization of education does not sacrifice global cognitive demands. Students are still evaluated using global standards, but within a context that is very familiar with their daily reality.

Structurally, the development of the instrument is mapped into two essential dimensions. The content dimension includes numbers, measurement and geometry, data and uncertainty, as well as algebra. Meanwhile, the cognitive dimension measures the hierarchy of understanding, application, and reasoning. From this matrix, 30 questions were formulated, integrated with 11 ethnomathematics stimuli from Javanese culture. The contexts covered include traditional food, dances, batik motifs, museum collections, and local community habits. The use of these local stimuli has proven to be pedagogically crucial. When students are faced with a cultural context they recognize, cognitive load will decrease drastically, allowing them to fully focus their mental capacity on mathematical reasoning, rather than on the confusion of translating an unfamiliar context.

Content Validity and Inter-Rater Reliability

Content validity was evaluated through expert judgment involving five validators from academia and primary education practice. The assessment covered three aspects: content relevance, language clarity, and presentation. Expert ratings were analyzed using Aiken's V coefficient, with items meeting the minimum validity threshold considered acceptable and those below the criterion requiring revision or removal. As shown in Table 1, all test stimuli achieved Aiken's V values of 0.80 or higher, indicating strong agreement among the validators and confirming the content validity of the instrument (Aiken, 1985b).

Table 1. Agreement validator coefficient

Stimulus	Aiken's V Index	Result	Stimulus	Aiken's V Index	Result
Traditional Snacks	0,95	Valid	Candi Singasari	0,95	Valid
Traditional Food Making Process	0,95	Valid	Mpu Tantular Museum's Collection	1	Valid
Remo Dance	1	Valid	East Java food	0,9	Valid
Batik Gajah Oling	0,95	Valid	Batik Industry	0,85	Valid
Wayang kulit	0,8	Valid	Karapan Sapi	0,95	Valid
Batik Kawung Pattern	0,9	Valid			

This phase of content validity testing was then followed by the measurement of assessment consistency. The level of agreement among the five experts was further evaluated using the Intraclass Correlation Coefficient (ICC) analysis, which proved that the five validators had uniform perceptions and measurement standards, ensuring that their assessment results were not influenced or coincidental. To ensure the accuracy of the computation, all statistical data processing for the ICC was accomplished using SPSS 24 software.

The Intraclass Correlation Coefficient (ICC) analysis yielded a reliability score of 0.909. To interpret this finding, we applied the evaluation benchmarks established by Hinton et al. (2004). Their framework classifies reliability into four distinct tiers: excellent (0.90 and above), high (0.70–0.90), moderate (0.50–0.70), and low (0.50 and below) (Hamed Taherdoost & Lumpur, 2016). Given that our coefficient clearly surpasses the 0.90 threshold, the inter-rater reliability of this instrument secures an excellent rating.

Item Validity and Reliability with Rasch Model

After filtering the initial draft of the numeracy instrument, we started a small-scale pilot study. The primary objective of this phase was to distinguish and remove any underlying defects such as uncertain wording or confusing layouts before the large-scale field deployment. For this introductory evaluation, a single cohort comprising 27 students from MIN 1 Surabaya City was selected. We managed the revised assessment directly to these participants, instructing them to complete the tasks within a strictly monitored time limit.

Following data collection, the empirical data were analyzed using the Rasch measurement model in Winsteps software with the Partial Credit Model (PCM), which accommodates items with multiple score categories. The analysis focused on evaluating both individual item characteristics and the instrument's overall psychometric quality by examining item fit, item difficulty, and summary statistics.

Item Fit Analysis

To separate the most valid questions, we implemented an item fit analysis. This diagnostic procedure is required to verify that each item functions predictably, confirming that accuracy measures the students' underlying numeracy skills without introducing statistical noise. We created the specific item diagnostics by accessing the "Table 10: Item Fit Order" output within Winsteps software.

The psychometric quality of the stimuli was arbitrated against three established statistical parameters. We relied heavily on the Outfit Mean-Square (MNSQ), the Outfit Z-Standard (ZSTD), and the Point Measure Correlation (Pt. Mean Corr.). If a specific question failed to meet the acceptable thresholds across these three indices, it was immediately flagged as psychometrically flawed (Bond & Fox, 2015). To protect the integrity of the measurement tool, we subsequently subjected these defective items to substantial revision or completely discarded them.

To measure whether each question functioned correctly, we adopted the fit criteria outlined by Sumintono and Widhiarso (2015a). According to their guidelines, an item is considered acceptable if it satisfies at least one of three statistical thresholds: an Outfit MNSQ between 0.5 and 1.5, an Outfit ZSTD between -2.0 and 2.0, or a Point Measure Correlation ranging from 0.4 to 0.85.

Following an evaluation of the 30-item instrument, 29 questions successfully met the requirements. However, item 16 exhibited severe misfit. Its diagnostic values entirely missed all three acceptable ranges (MNSQ = 1.70; ZSTD = 2.4; Pt. Mean Corr = -0.02). Because retaining such a defective question could skew the reliability of the overall measurement, we dropped item 16 and excluded it from all subsequent analyses.

Item Measure Analysis

Item difficulty was analyzed using Winsteps, which automatically ranks items from the most difficult to the easiest based on their logit measures reported in the Measure column of the Entry Number table. Difficulty levels were then classified according to the criteria proposed by Sumintono and Widhiarso (2015), which use the Rasch mean logit value (0.00) and standard deviation to categorize items as difficult, moderate, or easy.

The Item Measure output produced a standard deviation (SD) of 2.50. Using this rate alongside the 0.00 logit baseline, we separated the test items into three distinct difficulty tiers. Questions with logit scores exceeding +1 SD are characterized as highly challenging. Items falling within the -1 SD to +1 SD range signify a moderate difficulty level. Conversely, questions scoring below -1 SD are classified as easy for the students to answer (Adi et al., 2022). The whole breakdown of these item measures is initially summarized in Table 2.

Table 2 Difficulty Level based on Item Measure

Item	Measure	Result	Item	Measure	Result	Item	Measure	Result
4	3.82	Difficult	17	1.01	Moderate	13	-1.60	Moderate
3	3.31	Difficult	22	0.90	Moderate	12	-1.82	Moderate
11	3.04	Difficult	30	0.90	Moderate	20	-2.07	Moderate
5	2.95	Difficult	6	0.55	Moderate	25	-2.35	Moderate
16	2.77	Difficult	8	-0.19	Moderate	1	-2.70	Easy
14	2.67	Difficult	23	-0.86	Moderate	21	-2.70	Easy
9	2.47	Moderate	7	-0.91	Moderate	29	-2.70	Easy
26	2.47	Moderate	15	-1.04	Moderate	10	-3.16	Easy
24	2.00	Moderate	18	-1.21	Moderate	27	-5.14	Easy
28	1.37	Moderate	2	-1.60	Moderate	19	-5.33	Easy

As detailed further in Table 6, the 30 evaluated items are distributed across these three tiers. Six specific questions (items 4, 3, 11, 5, 16, and 14) fell into the difficult group. The bulk of the instrument—18 items in total—showed a moderate level of difficulty. The final six questions (items 1, 21, 29, 10, 27, and 19) made up the easy category.

Reliability

The following analysis is the instrument consistency analysis. The consistency analysis takes place with reliability assessment on the items deemed a fit. The output in Winsteps provides a reliability analysis based on the fit items. In Table 3, the reliability measured from 29 fit items shows a reliability score of 0.96. Based on that table, the instrument can be classified as having very high or exceptional consistency (Sumintono & Widhiarso, 2015b).

Table 3. Item Reliability

```
     SUMMARY OF 29 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      28.2       26.0         .00    .46      1.01     .0    .97    -.1 |
| S.D.      13.1         .0        2.36    .16       .31    1.5    .34    1.4 |
| MAX.      53.0       26.0        3.82   1.02      1.71    2.6   1.70    2.4 |
| MIN.       5.0       26.0       -5.33    .18       .20   -5.7    .22   -5.3 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .50 TRUE SD     2.31  SEPARATION  4.60  Item    RELIABILITY  .95 |
|MODEL RMSE    .49 TRUE SD     2.31  SEPARATION  4.77  Item    RELIABILITY  .96 |
| S.E. OF Item MEAN = .45                                                     |
-------------------------------------------------------------------------------
  MINIMUM EXTREME SCORE:      1 Item
UMEAN=.0000 USCALE=1.0000
```

Construct Validity Test of the Developed Numeracy Literacy Instrument

Construct validity was examined using Rasch analysis in Winsteps 3.73 to determine whether the instrument accurately measures the intended construct of numeracy literacy. The analysis was based on data from the large-scale field test, and construct validity was evaluated through the unidimensionality test. The unidimensionality results are presented in Figure 2.

```
    Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                            -- Empirical --    Modeled
Total raw variance in observations     =        53.7 100.0%        100.0%
  Raw variance explained by measures   =        24.7  46.0%         48.1%
    Raw variance explained by persons  =         5.1   9.5%         10.0%
    Raw Variance explained by items    =        19.6  36.5%         38.1%
  Raw unexplained variance (total)     =        29.0  54.0% 100.0%  51.9%
    Unexplned variance in 1st contrast =         3.0   5.6%  10.4%
    Unexplned variance in 2nd contrast =         2.1   4.0%   7.3%
    Unexplned variance in 3rd contrast =         2.0   3.8%   7.0%
    Unexplned variance in 4th contrast =         1.7   3.2%   5.8%
    Unexplned variance in 5th contrast =         1.5   2.8%   5.2%
```

Figure 2. Unidimensionality Test

The raw variance explained by measures shown in the figure above is 46.0% for empirical and 48.1% for modeled. According to Misbach and Sumintono (2014), the minimum requirement for unidimensionality is 20%, while Rasch unidimensionality requires at least 40%. Based on these data, the raw variance explained by measures meets the minimum requirement, falling into the good category (40–60%) (Prayoga et al., 2024).

Furthermore, an unexplained variance below 15% indicates good unidimensionality of the instrument (Adi et al., 2025; Noorkholisoh et al., 2024; Prayoga et al., 2024). The observed unexplained variances are 5.8%, 4.8%, 4.4%, 4.0%, and 3.5%, respectively. These results demonstrate that all unexplained variances remain below the 15% threshold, placing the instrument in the good category.

Table 4. Point Measure Correlation

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
16	42	156	2.74	.19	.93	-.7	.85	-1.2	.41	.31	76.3	73.9	i17	A
6	48	156	2.53	.18	1.00	.0	.97	-.3	.34	.33	72.4	71.2	i6	A
8	52	156	2.40	.18	1.02	.3	.92	-.8	.51	.33	78.8	69.8	i8	A
20	67	156	1.93	.17	.85	-2.7	.81	-2.7	.53	.34	76.9	65.9	i21	A
22	72	156	1.79	.17	.95	-.9	.93	-1.0	.41	.34	66.7	65.1	i23	A
15	87	156	1.35	.17	1.09	1.6	1.10	1.4	.22	.34	55.1	64.7	i15	A
17	95	156	1.11	.17	1.00	.0	.98	-.2	.34	.34	64.1	66.2	i18	A
28	99	156	.99	.18	.81	-2.9	.74	-3.1	.57	.33	70.5	67.3	i29	A
13	112	156	.57	.19	.92	-.9	.88	-.9	.47	.31	76.9	72.8	i13	A
2	119	156	.31	.20	.94	-.6	.92	-.5	.34	.30	78.2	76.5	i2	A
19	119	156	.31	.20	.95	-.5	.85	-1.0	.35	.30	76.9	76.5	i20	A
26	119	156	.31	.20	.83	-1.7	.72	-2.0	.49	.30	79.5	76.5	i27	A
10	125	156	.07	.21	1.07	.6	1.01	.1	.37	.28	76.9	80.0	i10	A
12	126	156	.03	.21	.96	-.3	.96	-.1	.27	.28	81.4	80.5	i12	A
3	129	156	-.01	.12	1.51	4.9	1.55	4.8	.38	.47	36.5	47.9	i3	B
4	130	156	-.02	.12	.62	-4.8	.68	-3.6	.33	.47	62.2	47.8	i4	B
5	147	156	-.26	.12	.81	-2.3	.83	-1.8	.30	.47	60.9	45.7	i5	B
14	176	156	-.65	.12	1.09	1.0	1.04	.5	.53	.47	42.3	48.7	i14	B
1	142	156	-.81	.26	.84	-.6	.88	-.3	.12	.22	91.0	89.4	i1	A
11	198	156	-.96	.12	.85	-1.7	.86	-1.4	.44	.46	53.2	51.1	i11	B
9	199	156	-.97	.12	.99	.0	.97	-.2	.42	.46	47.4	51.1	i9	B
21	204	156	-1.04	.12	.93	-.7	.87	-1.2	.59	.45	54.5	52.4	i22	B
24	146	156	-1.09	.27	.89	-.3	1.10	.4	-.05	.21	92.3	91.2	i25	A
25	211	156	-1.15	.12	1.05	.6	1.05	.5	.38	.44	51.3	53.8	i26	B
23	215	156	-1.21	.12	.97	-.3	.92	-.6	.44	.44	54.5	54.1	i24	B
27	219	156	-1.27	.13	1.18	1.8	1.14	1.2	.47	.44	54.5	55.0	i28	B
29	219	156	-1.27	.13	1.30	2.8	1.32	2.5	.42	.44	46.2	55.0	i30	B
18	277	156	-2.44	.17	1.26	1.6	1.14	.8	.28	.33	83.3	76.5	i19	B
7	358	156	-3.29	.09	1.37	2.8	2.34	3.5	.36	.54	47.4	50.2	i7	C
MEAN	146.6	156.0	.00	.16	1.00	-.1	1.01	-.3			65.8	64.7		
S.D.	70.8	.0	1.43	.04	.18	1.9	.30	1.8			14.9	13.0		

Construct validity in the Rasch model was evaluated using Item Polarity. Positive Point Measure Correlation (Pt. Mea-Corr) values across all items indicate consistency with the measured construct, while Outfit Mean Square (MNSQ) values below 1.5 reflect productive measurement (S. Hayati & Lailatussaadah, 2016). Positive Pt. Mea-Corr values show that items contribute appropriately to measuring the construct, whereas negative values indicate misfit and the need for item revision or removal.

As presented in Table 4, all items exhibit positive Pt. Mea-Corr values, indicating that they contribute effectively to measuring numeracy literacy. All Infit MNSQ values fall within the acceptable range (0.5–1.5), while only two items (i3 and i7) show Outfit MNSQ values slightly above the criterion, suggesting minor underfit likely caused by response inconsistency or random guessing. Because these values remain close to the threshold, both items were retained. Overall, the findings indicate a stable item difficulty hierarchy with positive item polarity, supporting the instrument's construct validity and its suitability for estimating numeracy literacy in this sample.

DISCUSSION

All items achieved Aiken's V values above 0.80, indicating that the instrument content accurately represents the cognitive construct of numeracy literacy through the alignment of test indicators, Javanese ethnomathematics contexts, and measurement objectives. As argued by Messick (1989), validity is established through accumulated evidence supporting score interpretation rather than being an inherent property of a test. Thus, the high level of expert agreement provides strong content-related validity evidence, consistent with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), which identify content quality as the primary foundation of assessment development.

The instrument also demonstrated excellent inter-rater reliability (ICC = 0.909), reflecting a high degree of agreement among validators and minimizing subjective judgment. Allen and Yen (2001) emphasize that such consistency is essential for supporting validity claims. Based on expert feedback, revisions were made to improve item wording, clarify instructions, and simplify several stimuli. These refinements follow the evidence-based test development principles of AERA et al. (2014) and help reduce measurement error so that students' scores more accurately reflect their mathematical reasoning.

Further evidence of construct validity was provided through Rasch analysis. Only one item failed to satisfy the item-fit criterion and was removed, while the remaining items fit the Rasch model, supporting accurate measurement and a stable internal structure. According to Bond and Fox (2015), fitting items produce more precise ability estimates, whereas removing misfitting items reduces measurement noise. In addition, the instrument demonstrated a balanced distribution of easy, moderate, and difficult items, enabling effective discrimination among students with different ability levels.

These findings are reinforced by the high item reliability index and satisfactory unidimensionality, two essential assumptions of Rasch Measurement Theory (Linacre, 2012; Sumintono & Widhiarso, 2015). Viewed through Messick's (1989) Unified Theory of Validity, the combined evidence from expert judgment, inter-rater agreement, and Rasch analysis forms a coherent body of validity evidence. Consequently, the instrument demonstrates strong psychometric quality while successfully integrating Javanese cultural contexts without compromising the measurement of numeracy literacy, making it suitable for use in Madrasah Ibtidaiyah classrooms.

CONCLUSION

This study confirms that the Javanese ethnomathematics-based numeracy instrument is psychometrically sound for Madrasah Ibtidaiyah students. Expert reviews indicated high content validity and strong agreement among raters. The Rasch analysis supported these initial findings, showing acceptable item fit, clear construct unidimensionality, and good overall reliability. Ultimately, the data suggest that this tool can measure students' numeracy skills consistently and fairly.

A key takeaway from the validation phase is the role of cultural integration. We found that incorporating Javanese ethnomathematical contexts into the test items did not interfere with the underlying numeracy construct. Instead, it created a more relatable testing environment that connects with the students' everyday backgrounds. Therefore, this assessment serves as a practical alternative for primary Islamic education, showing that local cultural values can be integrated into testing without lowering measurement standards.

One limitation to consider is that the current trial was restricted to a specific demographic in Surabaya City. Because of this localized sample, researchers should be cautious before generalizing these results. To build on this work, subsequent studies need to administer the assessment to a wider and more diverse range of students. We recommend running Differential Item Functioning (DIF) analyses to ensure the questions remain fair across different student subgroups. Looking ahead, examining how the instrument's scores correlate with other established academic variables will help secure its practical utility in real-world classrooms.

REFERENCES

Adi, N. R. M., Amaruddin, H., Adi, H. M. M., & A'yun, I. L. Q. (2022). Validity and reliability analysis using the rasch model to measure the quality of mathematics test items of vocational high schools. *Journal of Research and Educational Research Evaluation*, *11*(2), 103-113. https://journal.unnes.ac.id/sju/jere/article/view/58835

Adi, N. R. M. (2025). Analisis Validitas dan Reliabilitas Menggunakan Model Rasch untuk Mengukur Kualitas Instrumen Literasi Media. *Briliant: Jurnal Riset Dan Konseptual*, *10*(2), 323-336. https://doi.org/10.28926/briliant.v10i2.2048

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*.

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142.

Aini, Kurratul., Puspitorini, et.al. (2025). STEAM Learning: Integrating Deep Learning and Ethnomathematics to Improve Numeracy Literacy and Pancasila Student Profile.” *Al-Ishlah: Jurnal Pendidikan* 17, no. 4 (2025): 8040–52.

Allen, M. J., & Yen, W. M. (2001). *Introduction to Measurement Theory*.

Batiibwe, Kansiime, et.al. (2023). Adoption of Online Mathematics Learning in Ugandan Government Universities during the COVID-19 Pandemic: Pre-Service Teachers’ Behavioural Intention and Challenges. *Discover Education* 2, no. 1:12.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge. https://doi.org/10.4324/9781315814698

D’Ambrosio, U. (1985). Ethnomathematics and its Place in the History and Pedagogy of Mathematics. *For the Learning of Mathematics*, *5*(1).

Dewi, rahma. Kusuma, et.al. (2024). The Impact of The Independent Learning Curriculum’s Development on Mathematics Education. *Al Khawarizmi: Jurnal Pendidikan Dan Pembelajaran Matematika* 8, no. 2 (2024): 134–43.

Hamed Taherdoost, A., & Lumpur, K. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. In *International Journal of Academic Research in Management (IJARM)* (Vol. 5, Issue 3).

Hayati, S., & Lailatussaadah. (2016). Validitas dan Reliabilitas Instrumen Pengetahuan Pembelajaran Aktif, Kreatif, dan Menyenangkan (PAKEM) Menggunakan Model Rasch. *Jurnal Ilmiah DIDAKTIKA*, *16*(2), 169–179. https://doi.org/10.22373/jid.v16i2.593

Hidayat Ismail, H., Muhammad Amin Fauzi, K., & Sitompul, P. (2023). Perkembangan Penelitian Etnomatematika Di Indonesia. *Jurnal Pencerahan*, *17*(2), 1693–7775.

Hinton, P. R., Brownlow, C., Mcmurray, I. & Cozens, B. 2004. SPSS explained, East Sussex, England, Routledge Inc.

Linacre, J. M. (2012). *A User's Guide to Winsteps Rasch Model Computer Programs*. https://www.winsteps.com/winman/copyright.htm

Mahmudah, U., & Arif, S. (2022). Etnomatematika Sebagai Inovasi Pembelajaran dalam Mengintegrasikan Nilai Kearifan Lokal dan Konsep Matematika untuk Meningkatkan Hasil Belajar Siswa Madrasah Ibtidaiyah. *Cakrawala Jurnal Manajemen Pendidikan Islam Dan Studi Sosial*, *6*(2), 173–183. https://doi.org/10.33507/cakrawala.v6i2.1041

Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, *37*(18), 2734–2752. https://doi.org/10.1002/sim.7679

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-104). New York, NY: American Council on education and Macmillan.

Misbach, I. H., & Sumintono, B. (2014). Pengembangan dan validasi instrumen "Persepsi Siswa tehadap Karakter Moral Guru" di Indonesia dengan Model Rasch. *Seminar Nasional Psikometri*.

Noorkholisoh, L., Yustiana, Y. R., Budiman, N., & Suryana, D. (2024). Validity and Reliability Analysis Using the Rasch Model in Developing Creativity Tests Instruments for Elementary School Students. *Jurnal Ilmiah Bimbingan Konseling Undiksha*, *15*(1), 128–135. https://doi.org/10.23887/jibk.v15i1.81249

OECD. (2023a). PISA 2022 Results Factsheets Indonesia. *OECD (Organisation for Economic Co-Operation and Development) Publication*, 1–9.

OECD. (2023b). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Payadnya, I. P. A. A., Wulandari, I. G. A. P. A., Puspadewi, K. R., & Saelee, S. (2024). The significance of ethnomathematics learning: a cross-cultural perspectives between Indonesian and Thailand educators. *Journal for Multicultural Education*, *2020*.

Prahmana, R. C. I. (2022). Ethno-Realistic Mathematics Education: The Promising Learning Approach in the City of Culture." *SN Social Sciences* 2, no. 12 (2022): 257.

Prahmana, R. C. I., & D'Ambrosio, U. (2020). Learning geometry and values from patterns: Ethnomathematics on the batik patterns of yogyakarta, indonesia. *Journal on Mathematics Education*, *11*(3), 439–456.

Prayitno, Hadi S. et.al (2024). Unveiling The Reasoning Ability of Moderately-Skilled Elementary School Students in Learning Mathematics Through the Context of Eco-Mathematics." *Al-Bidayah: Jurnal Pendidikan Dasar Islam* 16, no. 1: 199–216.

Prayoga, K. P., Suryana, D., Supriatna, M., & Budiman, N. (2024). Penggunaan Rasch Model Untuk Menganalisis Konstruk Instrumen Kontrol Diri Pada Siswa Sekolah Menengah. *G-Couns: Jurnal Bimbingan Dan Konseling*, *9*(1), 367–381.

Rosa, Miltom, and Daniel Clark Orey (2011). Ethnomathematics: The Cultural Aspects of Mathematics." *Revista Latinoamericana de Etnomatemática Perspectivas Socioculturales de La Educación Matemática* 4, no. 2.

Safrida, L. N., Susanto, S., Suwito, A., & Ambarwati, R. (2024). Pengembangan Instrumen Literasi Numerasi Berbasis Etnomatematika Pertanian Bagi Siswa Sekolah Dasar. *EDU-MAT: Jurnal Pendidikan Matematika*, *12*(2), 383.

Sholihin, Fatoni K. et.al. (2024). Kemampuan Komunikasi Matematis Peserta Didik Pada Pembelajaran Matematika Terintegrasi Dengan Kebudayaan." *JEID: Journal of Educational Integration and Development* 4, no. 2 (2024): 66–79.

Sumintono, B., & Widhiarso, W. (2015a). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Penerbit Trim Komunikata.

Utari, T., Susanta, A., & Koto, I. (2024). Pengembangan Instrumen Asesmen Kompetensi Minimum (AKM) Numerasi Model Pisa Dengan Konteks Musi Rawas Kelas V Sekolah Dasar. *Jurnal Kapedas: Kajian Pendidikan Dasar*, *3*(1), 87–96.

Wirawan N, Syamsuri N,H, Fatah A (2024) Meta-analysis: ethnomathematics approaches in mathematical learning in Indonesia. World J Adv Res Rev 22(3):832-839. https://doi.org/10.30574/wjarr.2024.22.3.1778