MUBTADI: Jurnal Pendidikan Ibtidaiyah
Special Edition: The 1st ICEWISE
International Conference on Elementary Teaching and Wisdom in Education
https://ejournal.uinmadura.ac.id/index.php/ibtida
E-ISSN: 2720-8850 P-ISSN: 2715-7067

DEVELOPMENT OF A NORTH SUMATRAN LOCAL CULTURE BASED NUMERACY ASSESSMENT INSTRUMENT FOR ELEMENTARY SCHOOL STUDENTS

Asriana Harahap[1], Ira Aniati[2], Mhd. Latip Kahpi[3], Samsuddin Pulungan[4], Fajar Dwi Mukti[5]
Universitas Islam Negeri Syekh Ali Hasan Ahmad Addary Padangsidimpuan[1,2,3,4], Dokuz Eylul University Turkey[5]
asrianaharahap@uinsyahada.ac.id[1], iraaniati@uinsyahada.ac.id[2], latipkahpi@uinsyahada.ac.id[3], samsuddin@uinsyahada.ac.id[4], 2025950356@ogr.deu.edu.tr[5]

Abstrak
Penelitian ini dilatarbelakangi oleh kebutuhan instrumen asesmen numerasi yang valid, reliabel, dan kontekstual dengan keragaman budaya lokal Sumatera Utara. Penelitian ini bertujuan mengembangkan dan menguji kualitas instrumen asesmen numerasi berbasis budaya lokal bagi siswa sekolah dasar. Penelitian menggunakan metode Research and Development (R&D) dengan model 4D yang meliputi tahap Define, Design, Develop, dan Disseminate. Subjek penelitian adalah siswa kelas V sekolah dasar dari tiga kabupaten/kota di Sumatera Utara, dengan 15 siswa pada uji coba skala kecil dan 96 siswa pada uji coba lapangan utama yang dipilih melalui teknik purposive sampling. Data dikumpulkan melalui lembar validasi ahli materi, evaluasi, dan budaya serta tes numerasi yang mengintegrasikan konteks budaya Batak Toba, Karo, Simalungun dan Mandailing. Hasil penelitian menunjukkan bahwa instrumen memperoleh validitas ahli yang sangat tinggi, dengan skor rata-rata 90,9% pada aspek materi, 87,3% pada aspek konstruksi, 88,6% pada aspek bahasa, dan 93,6% pada aspek budaya. Koefisien reliabilitas komposit mencapai 0,84 atau kategori tinggi. Dari 27 butir soal, sebanyak 24 butir memenuhi kriteria validitas empiris, sedangkan tiga butir lainnya masih memerlukan revisi atau eliminasi. Dengan demikian, instrumen yang dikembangkan menunjukkan potensi alternatif asesmen numerasi berbasis budaya lokal, tetapi kelayakan keseluruhan instrumen masih memerlukan penyempurnaan terhadap ketiga butir tersebut dan pengujian lebih lanjut.
Kata Kunci: Asesmen Numerasi, Budaya Lokal, Instrumen Penilaian, Sekolah Dasar, Sumatera Utara

Abstract
This study addresses the need for valid, reliable, and culturally relevant numeracy assessment instruments for elementary school students in North Sumatra. It aimed to develop and evaluate a numeracy assessment instrument incorporating local cultural contexts. The study employed a Research and Development (R&D) approach using the 4D model: Define,

Design, Develop, and Disseminate. Participants were fifth-grade students from three regencies/cities in North Sumatra, selected through purposive sampling. The small-scale trial involved 15 students, while the main field trial included 96 students. Data were collected through validation by content, assessment, and cultural experts and numeracy tests integrating Batak Toba, Karo, Simalungun, and Mandailing cultural contexts. Expert validation yielded average scores of 90.9% for content, 87.3% for item construction, 88.6% for language, and 93.6% for cultural relevance, indicating very high validity. The composite reliability coefficient was 0.84, indicating high reliability. Of the 27 items, 24 met empirical validity criteria, while three required revision or elimination. These findings suggest that the developed instrument offers a promising alternative for culturally contextualized numeracy assessment. Nevertheless, the instrument requires refinement of the three problematic items and further testing before its overall suitability can be established. Further validation with broader samples is recommended to support future generalizability.

Keywords: Numeracy Assessment; Local Culture; Assessment Instrument; Elementary School; North Sumatra

Received: 17-10-2026	Accepted: 09-11-2026	Published: 23-11-2026

©Mubtadi: Jurnal Pendidikan Ibtidaiyah
Prodi Pendidikan Guru Madrasah Ibtidaiyah
Universitas Islam Negeri Madura, Indonesia

https://doi.org/10.19105/mubtadi.v1i.26414

INTRODUCTION

Primary education serves as a fundamental foundation for developing high-quality human resources. Within this context, mathematics education, particularly numeracy, plays an essential role in preparing students to address academic challenges and practical demands in everyday life (Citrawan et al., 2024). In an increasingly complex and data-driven society, numeracy is no longer understood merely as the ability to perform mathematical calculations. Rather, it encompasses the capacity to interpret quantitative information, apply mathematical concepts, and make informed decisions in various real-life situations (Richards & Islam, 2018). Numeracy competence is therefore a fundamental skill that supports students' academic development and their readiness to participate effectively in contemporary society (Essiet et al., 2023). It involves the ability to mobilize mathematical knowledge, facts, and tools to solve problems in meaningful everyday contexts (Leite, 2024).

Despite its importance, Indonesian students' numeracy achievement remains a significant educational concern. The results of the Programme for International Student Assessment (PISA) 2022 indicate that Indonesian students' mathematical performance remains below the average of Organisation for Economic Co-operation and Development (OECD) participating countries (Hortsch et al., 2026). Similarly, findings from the Minimum Competency Assessment (Asesmen Kompetensi Minimumor AKM) indicate that elementary school students continue to demonstrate

varying levels of numeracy proficiency, with some students requiring additional support to achieve basic competency standards (Kelana et al., 2020). These conditions highlight the importance of strengthening not only numeracy instruction but also the quality of assessment instruments used to identify and measure students' mathematical competencies (Suherman & Vidákovich, 2022) .

The Minimum Competency Assessment (AKM) provides an important framework for understanding the dimensions of numeracy assessment in Indonesia. Numeracy in the AKM framework encompasses three interconnected dimensions: content, cognitive processes, and contexts (Eraikhuemen & Aikpitanyi, 2015a) . The content dimension includes numbers, algebra, geometry and measurement, as well as data and uncertainty. The cognitive process dimension comprises understanding, application, and reasoning, while the contextual dimension includes personal, sociocultural, and scientific contexts (Fu et al., 2026). The inclusion of sociocultural contexts demonstrates that cultural backgrounds are recognized as legitimate components of national numeracy assessment. Consequently, developing local culture-based numeracy assessment instruments does not represent a departure from the AKM framework. Instead, it provides an opportunity to strengthen the sociocultural dimension, particularly by incorporating cultural characteristics that have not been systematically optimized in the assessment context of North Sumatra (No et al., 2012) .

The integration of local cultural contexts into numeracy assessment is also relevant to the cognitive developmental characteristics of elementary school students. According to Piaget's theory of cognitive development, children at the concrete operational stage generally rely on concrete objects, observable experiences, and familiar representations to understand abstract concepts (Sun & Buripakdi, 2024). Therefore, presenting mathematical problems through familiar cultural objects and everyday practices can help students establish meaningful connections between concrete experiences and mathematical concepts (Bellini et al., 2019; Eraikhuemen & Aikpitanyi, 2015b) . For example, geometric patterns found in traditional textiles, architectural structures, and cultural artifacts can provide contextual representations for understanding mathematical relationships. This approach is consistent with culturally responsive teaching, which emphasizes the integration of students' cultural backgrounds into educational practices to promote meaningful learning experiences (Wang et al., 2026).

One factor that may contribute to students' difficulties in numeracy assessment is the limited connection between assessment contexts and their everyday cultural experiences. Numeracy instruments that predominantly employ generic or culturally neutral contexts may not fully accommodate the diverse backgrounds of students, particularly those living in communities with distinctive cultural traditions (Barumbun et al., 2023). In this regard, ethnomathematics offers an important perspective by examining mathematical ideas, practices, and representations embedded in cultural activities and artifacts. Previous studies have demonstrated that integrating cultural contexts into mathematics education can support students' mathematical reasoning, conceptual understanding, learning motivation, and engagement (Niklas et al., 2022). Nevertheless, the educational potential of ethnomathematics needs to be translated beyond the identification of mathematical concepts in cultural artifacts into the systematic development of assessment instruments that can measure students' actual numeracy competencies (Karatepe Mumcu, 2026) .

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

North Sumatra is characterized by considerable cultural diversity, including the Batak Toba, Karo, Simalungun, Pakpak, Mandailing, Coastal Malay, and Nias communities. This diversity provides substantial opportunities for integrating local cultural knowledge into mathematics education and assessment (Ali et al., 2025). Several cultural elements contain mathematical concepts that can be developed into meaningful numeracy contexts. For instance, geometric patterns in Batak Toba ulos textiles provide representations of mathematical structures and spatial relationships (Arifa et al., 2023). The Siwaluh Jabu traditional house of the Karo community offers architectural elements that can be connected to geometric concepts and spatial measurement. Similarly, Bagas Godang, the traditional house of the Mandailing community, provides cultural and architectural contexts for exploring mathematical relationships. These cultural elements demonstrate the potential of local knowledge as a source of contextualized mathematical problems. However, previous ethnomathematical investigations have predominantly emphasized identifying and describing mathematical concepts within cultural artifacts, while their transformation into systematically developed numeracy assessment instruments remains an important area for further investigation (Munadi & Ratna Febriyanti, 2020).

The development of local culture-based numeracy assessment instruments has also been explored in several regions of Indonesia, including studies drawing on Minangkabau, Malay, Banten, and South Kalimantan cultural contexts (Bamber, 2019). These studies provide valuable foundations for connecting mathematical assessment with students' cultural environments. However, the main challenge is not simply the limited number of studies addressing local culture-based numeracy instruments (Ali et al., 2025). More importantly, it concerns the extent to which previously developed instruments have been systematically examined in terms of their psychometric quality, particularly content validity and empirical item analysis (de Bruyn et al., 2026) . An instrument may receive positive evaluations from experts or demonstrate satisfactory overall reliability, yet individual items may still fail to meet empirical validity criteria (Starmer et al., 2010) . Therefore, examining item-level performance is essential to determine whether each assessment item adequately represents the intended numeracy competencies and functions appropriately when administered to students (Zulaeha & Sangadji, 2025) .

This issue becomes particularly important in developing assessment instruments for multi-ethnic cultural contexts. Integrating several cultural backgrounds into a single instrument requires more than incorporating cultural names, objects, or traditional representations into mathematical questions (Su et al., 2026). Each cultural context must be appropriately represented, educationally relevant, and aligned with the intended numeracy competencies. Furthermore, the instrument needs to undergo systematic content validation involving relevant experts and empirical item analysis to identify items that require improvement or elimination. Accordingly, the research gap addressed in this study concerns the need for a local culture-based numeracy assessment instrument for elementary school students that integrates multiple cultural contexts in North Sumatra and examines its quality through both expert validation and item-level empirical analysis. This focus extends previous ethnomathematical investigations by emphasizing not only cultural representation but also the assessment instrument's measurable quality and practical potential (Cheruiyot & Molnár, 2026).

Based on this rationale, the present study aims to develop and evaluate a local culture-based numeracy assessment instrument for fifth-grade elementary school students in North Sumatra. The instrument integrates cultural contexts from three ethnic communities: Batak Toba, Karo, and Mandailing. Its development refers to relevant numeracy content domains within the AKM framework, particularly numbers, geometry and measurement, and data and uncertainty. Using a Research and Development (R&D) approach with the 4D model (Define, Design, Develop, and Disseminate), this study examines the instrument's quality through expert validation covering content, construction, language, and cultural relevance, as well as empirical testing of individual items and overall reliability.

The novelty of this study lies in three main contributions. First, it integrates the cultural contexts of Batak Toba, Karo, and Mandailing communities into a single numeracy assessment instrument, extending the focus of previous ethnomathematical studies that have generally examined individual cultural contexts (Calvo et al., 2014). Second, it specifically addresses fifth-grade elementary school students, providing an assessment instrument grounded in cultural contexts relevant to the developmental characteristics of primary education learners (Dewanti et al., 2025). Third, it evaluates instrument quality through multiple dimensions of expert validation, composite reliability, and empirical item-level analysis, thereby identifying both items that meet validity criteria and those requiring revision or elimination. This item-level evaluation provides a more comprehensive understanding of the instrument's strengths and limitations than relying solely on overall validation and reliability scores (Urwick, 2022).

Beyond its methodological contribution, this study has implications for educational equity and culturally responsive assessment practices. North Sumatra encompasses diverse geographical and cultural environments, ranging from urban communities to highland and island regions. Differences in students' cultural experiences may influence how they interpret and respond to assessment questions. Integrating familiar cultural contexts into numeracy assessment is therefore expected to support more meaningful assessment experiences and reduce potential barriers arising from unfamiliar problem contexts. Ultimately, this study seeks to contribute to the development of more contextualized numeracy assessment practices while providing an empirical foundation for further refinement and evaluation of culturally responsive assessment instruments in Indonesian elementary education.

METHODS

Research Design and Development Procedure

This study employed a Research and Development (R&D) approach using the 4D development model, consisting of four stages: Define, Design, Develop, and Disseminate (Randy et al., 2024). The model was selected because it provides a systematic procedure for developing an educational assessment instrument, beginning with needs identification and ending with product evaluation and dissemination.

The Define stage involved three operational activities. First, curriculum analysis was conducted to identify the numeracy competencies expected of fifth-grade elementary school students, particularly those associated with the Minimum Competency Assessment (AKM)

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

framework. Second, a needs analysis was conducted through discussions and information gathering involving elementary school teachers and students to identify difficulties in numeracy assessment and the relevance of familiar cultural contexts. Third, an inventory of local cultural elements was compiled to identify cultural objects, traditional architecture, textile patterns, and community economic activities that could be transformed into mathematical assessment contexts. The selected cultural elements were limited to those relevant to the Batak Toba, Karo, and Mandailing communities.

The Design stage involved preparing the instrument blueprint and constructing the initial assessment items. The blueprint connected numeracy content domains, learning indicators, cognitive demands, item formats, and local cultural contexts. Based on this blueprint, 27 assessment items were developed, consisting of 20 multiple-choice questions, four short-answer questions, and three open-ended questions. The initial items were reviewed internally to ensure alignment between the questions, assessment indicators, mathematical content, and cultural representations.

The Develop stage consisted of expert validation, instrument revision, a small-scale trial, and a main field trial. Five experts evaluated the instrument, comprising two experts in numeracy content and assessment, one language expert, and two local cultural experts. Revisions were made based on their recommendations before the instrument was administered to 15 fifth-grade students in the small-scale trial. The revised instrument was subsequently administered to 96 fifth-grade students in the main field trial. The trial results were analyzed to examine empirical item validity, item difficulty, item discrimination, and reliability.

The Disseminate stage was conducted on a limited scale by communicating the development process and research findings to the elementary school teachers involved as research partners. The dissemination activities were intended to provide information about the instrument's design, cultural contexts, assessment procedures, and evaluation results. This stage did not constitute a large-scale implementation or a national effectiveness evaluation.

Research Subjects and Participant Characteristics

The research subjects comprised three groups: expert validators, small-scale trial participants, and main field trial participants. The expert validation involved five experts selected based on their academic expertise and relevant experience. Two experts assessed numeracy content and assessment construction, one expert evaluated language clarity and appropriateness, and two experts assessed the accuracy and relevance of the cultural representations incorporated into the instrument. The small scale trial involved 15 fifth grade elementary school students. This trial was conducted to examine the clarity of instructions, comprehensibility of questions, appropriateness of cultural stimuli, and practical administration of the instrument before the main field trial.

The main field trial involved 128 fifth grade elementary school students from three elementary schools in North Sumatra, with 32 students recruited from each participating school. Participants were selected through purposive sampling based on the schools' geographical and cultural characteristics and their relevance to the cultural contexts represented in the assessment instrument. The selection was intended to provide variation in students' educational and sociocultural environments.

The participants' characteristics were documented according to school location, geographical setting, grade level, and number of students. The distribution of participants is presented in Table 1.

Table 1 - Characteristics of Participants in the Main Field Trial

School	Regency/City	Regional Characteristics	Number of Students
SDN 1 Padangsidimpuan	Padangsidimpuan City	Highland area, located near the urban region of Padangsidimpuan	32
SDN 1 Balige	Toba Regency	Highland area, located near the Lake Toba region	32
SDN 1 Karo	Karo Regency	Highland area, agricultural center	32
SDN 1 Simalungun	Simalungun Regency	Highland area, agricultural center	32

The study adhered to research ethics principles. Research permission was obtained from the participating schools and relevant education authorities. The research objectives and procedures were explained to teachers and parents or guardians, and informed consent was obtained before students participated. Students' identities were protected through the use of identification codes during data processing and reporting.

Instrument Development and Blueprint

The primary research instrument was a local culture-based numeracy assessment test consisting of 27 items. The instrument was developed with reference to three numeracy content domains within the AKM framework: numbers, geometry and measurement, and data and uncertainty. The assessment incorporated cultural contexts from the Batak Toba, Karo, Simalungun and Mandailing communities. These contexts were selected based on their relevance to mathematical concepts, familiarity within the target communities, and suitability for fifth-grade students. Examples included ulos textile motifs, traditional architectural structures such as Rumah Bolon, Siwaluh Jabu, and Bagas Godang, and traditional economic activities and local community practices. The instrument blueprint was structured as follows.

Table 2 - Blueprint of the Local Culture-Based Numeracy Assessment Instrument in North Sumatra

Content Domain	Indicator	Cultural Context	Number of Items
Numbers	Comparing and performing operations involving whole numbers and simple fractions	Traditional market transactions; harvest products such as frankincense and coffee	9
Geometry and Measurement	Identifying two-dimensional shapes and measuring	Ulos motifs; roof structures of Rumah Bolon, Siwaluh Jabu, and Bagas Godang	10

	length/area using non-standard units		
Data and Uncertainty	Reading, presenting, and interpreting simple data	Number of Gordang Sambilan musical instruments; agricultural production data among villages	8

Each item was constructed according to the following principles: (1) alignment with the specified numeracy content and assessment indicators; (2) clarity and accuracy of mathematical information; (3) appropriateness of language for fifth-grade students; (4) accuracy of cultural representations; and (5) suitability of the cognitive demands for the target students.

Data Collection Procedures

Data were collected through expert validation and student testing. The expert validation process was conducted before the field trials. Each validator received the instrument draft and an assessment sheet containing indicators relevant to their area of expertise. Validators evaluated the items independently using a four-point Likert scale.

The four assessment categories were: 1 = not appropriate, 2 = less appropriate, 3 = appropriate, and 4 = highly appropriate. The evaluated aspects included content relevance, item construction, language clarity, and cultural appropriateness. Written comments and recommendations were also collected to identify items requiring revision.

Following expert validation and revision, the instrument was administered to 15 students in the small-scale trial. Feedback from this trial was used to improve the clarity of instructions, wording, and presentation of assessment items. The revised instrument was subsequently administered to 96 students in the main field trial under standardized administration conditions. Students received the same instructions, working time, and assessment procedures. Completed responses were collected, coded, and entered into the analysis dataset.

Scoring Procedures

The assessment instrument consisted of three item formats: multiple-choice, short-answer, and open-ended questions. Each response was evaluated according to a predetermined scoring procedure to ensure consistency in the assessment process.

Multiple-choice items were scored dichotomously, with a score of 1 for a correct answer and 0 for an incorrect answer or an unanswered item. Short-answer items were evaluated according to the accuracy and completeness of the mathematical response. Open-ended items were assessed using an analytical scoring rubric that considered the correctness of the mathematical approach, the accuracy of calculations, and the appropriateness of the final answer.

For short-answer and open-ended questions, the scoring rubric was prepared before the field trial and applied consistently to all participants. The maximum score for each item was determined according to its scoring rubric. The total score for each student was obtained by summing the scores across all 27 items.

Content Validity and Instrument Evaluation Criteria

Content validity was evaluated using Aiken's V coefficient to determine the extent to which each assessment item represented the intended content and assessment indicators (Aiken, 1985).

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

With five validators and a four-point rating scale, each item received an Aiken's V coefficient between 0 and 1. The coefficients were examined individually to identify items requiring revision. For this study, the following operational interpretation was proposed: coefficients of 0.80–1.00 indicated high content validity; 0.60–0.79 indicated that an item required review or revision; and coefficients below 0.60 indicated insufficient content validity and the need for substantial revision or elimination. These categories should be interpreted as the study's evaluation criteria rather than universal statistical cutoffs.

In addition to Aiken's V, expert validation percentages were calculated to summarize the overall evaluations of the four instrument aspects: content, construction, language, and cultural appropriateness. The results were interpreted using the validity classification adopted by the study. Item-level Aiken's V results and aspect-level validation percentages were reported separately to distinguish individual item validity from the overall expert assessment.

Empirical Item Analysis and Reliability

The main field trial data were analyzed to evaluate the empirical quality of the 27 assessment items. The analysis included item validity, item difficulty, item discrimination, and reliability. Item validity, The empirical validity of each item was examined using the point-biserial correlation for dichotomously scored multiple-choice items. For short-answer and open-ended items, corrected item-total correlation or an appropriate product-moment correlation was used, depending on the scoring characteristics of the item. An item was considered empirically valid when its correlation coefficient was positive and statistically significant at the 5% significance level. For the main field trial involving 96 students, the correlation results were evaluated against the appropriate critical value for the sample size and significance level.

Item difficulty, Item difficulty was examined to determine the proportion of students who answered an item correctly. For dichotomous items, the difficulty index was calculated using: Higher values indicate easier items, whereas lower values indicate more difficult items. The conventional interpretation used was: 0.00–0.30 = difficult, 0.31–0.70 = moderate, and 0.71–1.00 = easy. For polytomous items, the proportion of the mean item score to the maximum possible item score was used as an adjusted difficulty index.

Item discrimination, Item discrimination was examined to determine the extent to which an item differentiated between students with higher and lower numeracy performance. The discrimination index was calculated using the difference between the proportion of correct responses in the upper and lower performance groups. The conventional interpretation was: 0.40 or above = very good; 0.30–0.39 = good; 0.20–0.29 = requires revision; and below 0.20 = poor. For items with polytomous scores, item-total relationships were also examined to identify items that did not adequately distinguish between different levels of student performance.

Reliability analysis, Internal consistency was examined using the Kuder-Richardson Formula 20 (KR-20) for dichotomously scored items and Cronbach's Alpha for polytomously scored items. The coefficient was interpreted in relation to the consistency of student responses across assessment items. A reliability coefficient of 0.70 or higher was considered acceptable for exploratory instrument evaluation, while coefficients approaching or exceeding 0.80 indicated good internal consistency. Where a composite reliability coefficient was reported, the calculation

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

procedure and underlying statistical model were specified separately to avoid confusing composite reliability with KR-20 or Cronbach's Alpha.

The results of the empirical analyses were used to classify each item into three categories: (1) items meeting the validity and quality criteria and retained in the instrument; (2) items requiring revision based on their empirical and substantive characteristics; and (3) items recommended for elimination when their deficiencies could not be adequately addressed through revision. The final instrument evaluation considered expert judgments, item-level statistical results, and the overall reliability evidence rather than relying on a single statistical coefficient.

All quantitative data were processed using descriptive statistical analysis and appropriate psychometric procedures. The results were presented in tables containing expert validation scores, item validity coefficients, difficulty indices, discrimination indices, and reliability coefficients to ensure transparency and support replication of the study.

RESULTS AND DISCUSSION

Results of Instrument Development

The development of the North Sumatran Local Culture-Based Numeracy Assessment Instrument followed the four stages of the 4D development model: Define, Design, Develop, and Disseminate. Each stage was conducted to ensure that the instrument addressed students' numeracy learning needs while incorporating culturally relevant contexts.

Define Stage: Needs Analysis and Identification of Cultural Contexts

The needs analysis revealed that 82% of the participating teachers had never used numeracy assessment instruments that explicitly integrated local cultural contexts. This finding indicates a gap between the potential of culturally contextualized learning resources and the assessment practices commonly implemented in elementary schools. Teachers generally relied on numeracy questions using general or nationally standardized contexts, which did not always reflect students' immediate sociocultural environments.

Students' responses also indicated that they were more familiar with local cultural symbols, objects, and everyday activities than with some of the general contexts commonly presented in Minimum Competency Assessment (AKM) items. This finding provided an initial rationale for developing an assessment instrument that connects mathematical concepts with students' cultural experiences.

Three cultural contexts were selected for the development process: Batak Toba, Karo, and Mandailing. The selection was based on their relevance to the geographical and sociocultural settings of the participating schools and their potential to represent numeracy concepts through familiar objects, traditional architecture, textile patterns, economic activities, and cultural performances.

Design Stage: Development of the Assessment Blueprint

At the Design stage, the numeracy content domains were mapped into relevant cultural contexts. The blueprint covered three main domains: Numbers; Geometry and Measurement; and Data and Uncertainty. Cultural contexts were selected to ensure that the assessment items measured mathematical competencies rather than merely students' knowledge of cultural terminology.

The blueprint was subsequently developed into an initial draft containing 30 assessment items. The items consisted of multiple-choice questions, short-answer questions, and open-ended questions. Each item was designed to assess a specific numeracy competency while presenting a contextual problem related to the selected cultural settings.

Develop Stage: Expert Validation and Instrument Revision

The initial draft was evaluated by five experts with relevant expertise in numeracy content and assessment, language, and local culture. Expert validation focused on four aspects: content relevance, item construction, language clarity, and cultural appropriateness. During the Develop stage, the draft instrument was validated by five experts. The summary of expert validation results is presented in Table 3.

Table 3 - Summary of Expert Validation Results

Validation Aspect	Mean Score (%)	Category
Content	90.9	Highly Valid
Construction	87.3	Highly Valid
Language	88.6	Highly Valid
Cultural Appropriateness	93.6	Highly Valid

As shown in Table 3, all four aspects received high validation scores. Cultural appropriateness obtained the highest score (93.6%), suggesting that the experts considered the selected cultural representations relevant and appropriate for the intended assessment context. The construction aspect received the lowest score (87.3%), indicating that some technical features of the questions, including the clarity of illustrations and the formulation of assessment prompts, required particular attention.

Following the experts' recommendations, the initial 30-item draft underwent revisions. Three items were revised because their visual illustrations or representations were unclear, while two items were eliminated because their cultural contexts were considered insufficiently familiar or appropriate for the intended student population. Consequently, the instrument entering the subsequent field-testing stage consisted of 27 items. The revision process was intended not only to improve the technical quality of the questions but also to ensure that cultural representations supported the measurement of numeracy competencies rather than introducing unnecessary interpretation difficulties.

Table 4 - Results of Item Analysis and Instrument Reliability

Psychometric Indicator	Result	Interpretation
Number of valid items	24 out of 27 items (88.9%)	Good
KR-20 Reliability (dichotomous items)	0.81	High
Cronbach’s Alpha Reliability (open-ended items)	0.78	High
Composite Reliability	0.84	High
Item difficulty distribution (easy:moderate:difficult)	26% : 52% : 22%	Proportional
Mean item discrimination index	0.38	Good

The empirical analysis of the 27-item instrument indicated that 24 items met the established item-validity criteria, representing 88.9% of the tested items. The remaining three items had item-

validity coefficients below 0.20 and therefore did not meet the minimum criterion applied in this study. These items were associated with Karo cultural contexts that may have been less familiar to some students outside Karo Regency.

The composite reliability coefficient of 0.84 indicates that the instrument demonstrated a high level of measurement consistency. The distribution of item difficulty levels also showed a relatively balanced proportion among easy, moderate, and difficult categories. Meanwhile, the mean item discrimination index of 0.38 was categorized as good, indicating that the instrument was adequately capable of distinguishing between students with higher and lower numeracy abilities.

The results of the student response questionnaire after the trial also revealed that 87% of students reported feeling interested and familiar with the cultural contexts incorporated into the assessment items. As a concrete illustration, Table 5 presents several examples of developed assessment items along with their cultural contexts and measured numeracy domains.

Table 5 - Examples of Assessment Items Based on Cultural Contexts

Domain	Cultural Context	Description of Assessment Item
Numbers	Karo traditional market	Calculating the total price of agricultural products (cabbage and carrots) sold by kilogram
Geometry and Measurement	Batak Toba ulos motifs	Identifying two-dimensional geometric shapes (triangles and squares) found in ulos textile patterns
Geometry and Measurement	Siwaluh Jabu traditional house	Measuring the comparison of traditional house pillar heights using non-standard measurement units
Data and Uncertainty	Gordang Sambilan	Reading a bar chart showing the number of performers in Gordang Sambilan performances across five villages
Numbers	Mandailing Bagas Godang	Calculating the number of stairs and spatial divisions within Bagas Godang

In addition to overall item analysis, further analysis was conducted based on each numeracy content domain, as presented in Table 6.

Table 6 - Item Analysis Based on Numeracy Content Domains

Content Domain	Number of Items	Mean Difficulty Level	Mean Discrimination Index	Reliability Category
Numbers	9	0.54 (Moderate)	0.36	High
Geometry and Measurement	10	0.58 (Moderate)	0.41	High
Data and Uncertainty	8	0.49 (Moderate)	0.35	High

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

Based on Table 6, all three numeracy domains demonstrated moderate average difficulty levels and good discrimination indices. The geometry and measurement domain showed the highest discrimination index (0.41). This finding indicates that assessment items connecting geometric concepts with concrete cultural objects, such as traditional houses and textile motifs, were relatively more effective in distinguishing between high- and low-performing students compared with items in the data and uncertainty domain.

The relatively lower discrimination index in the data and uncertainty domain may be attributed to the higher level of abstraction required when interpreting diagrams and graphical representations, which remains challenging for fifth-grade elementary school students.

The results indicate that the three numeracy domains demonstrated moderate average difficulty levels. Geometry and Measurement obtained the highest mean discrimination index (0.41), followed by Numbers (0.36) and Data and Uncertainty (0.35). This pattern suggests that the geometry-related items showed relatively stronger differentiation between students with different levels of numeracy performance in the field trial.

Nevertheless, these domain-level averages should be interpreted alongside the item-level results. An average discrimination index does not establish that every item within a domain performs adequately. Similarly, domain-level reliability estimates should only be reported if they have been calculated separately from the relevant item-level response data.

Discussion

Psychometric Quality of the Developed Instrument

The findings demonstrate that the development of a local culture-based numeracy assessment instrument requires attention to both content validity and empirical item performance. The expert validation results indicate that the instrument has strong content relevance, appropriate construction, clear language, and cultural suitability. The high score for cultural appropriateness (93.6%) is particularly important because cultural integration must be based on accurate representations of local traditions and meaningful connections with the numeracy competencies being assessed.

However, expert validation alone cannot establish the empirical quality of an assessment instrument. The field-trial results provide an additional basis for evaluating whether individual items function adequately when administered to students. In this study, 24 of the 27 items met the empirical validity criterion, while three items did not. This distinction is essential: the instrument demonstrated promising overall psychometric characteristics, but its empirical quality was not uniform across all items.

The mean discrimination index of 0.38 also suggests that the instrument was generally capable of distinguishing students with different levels of numeracy performance. The moderate distribution of item difficulty indicates that the instrument included questions requiring different levels of student competence. Such variation is important in educational assessment because an instrument consisting exclusively of easy or difficult questions may provide limited information about students' actual numeracy abilities.

Nevertheless, the three items that failed to meet the empirical validity criterion require further investigation. Their low coefficients may reflect problems in item construction, cultural familiarity, visual presentation, or the relationship between the cultural context and the mathematical competency being measured. Therefore, these items should be revised or eliminated based on further item-level examination and subsequently tested again.

The reliability findings should also be interpreted according to the scoring structure. The KR-20 coefficient of 0.81 for multiple-choice items and Cronbach's alpha coefficient of 0.78 for constructed-response items indicate acceptable internal consistency within their respective item groups. The reported composite reliability of 0.84 provides further evidence of measurement consistency, provided that its calculation procedure is clearly documented. These findings support the potential use of the instrument while emphasizing that the three problematic items must be addressed before the complete instrument can be considered ready for broader implementation.

Cultural Diversity and Fairness in Numeracy Assessment

One of the main contributions of this study is its attempt to incorporate the cultural diversity of North Sumatra into numeracy assessment. The use of Batak Toba, Karo, and Mandailing contexts provides opportunities to connect mathematical concepts with objects, practices, and cultural representations that are part of students' social environments.

The integration of local culture can make numeracy problems more meaningful by presenting mathematical concepts through familiar situations. For example, traditional market transactions can be used to assess numerical operations, ulos motifs can introduce geometric concepts, traditional houses can provide contexts for measurement, and cultural performance data can be used to assess students' ability to interpret graphs.

The integration of these cultural contexts is relevant to culturally responsive assessment because it connects mathematical knowledge with students' sociocultural experiences. However, cultural contextualization does not automatically guarantee assessment fairness. A cultural reference that is familiar to one group of students may be less familiar to students from another ethnic or geographical background.

The finding that three items containing Karo cultural contexts did not meet the empirical validity criterion illustrates this issue. It raises the possibility that differences in cultural familiarity may influence how students interpret assessment questions. Nevertheless, the present findings do not establish that cultural background caused the low item-validity coefficients. Such a conclusion would require additional evidence, including student interviews, cognitive interviews, and statistical analyses of potential item bias.

Therefore, cultural diversity should be considered not merely as a source of contextual material but also as an important factor in assessment design and evaluation. Assessment developers need to ensure that cultural references provide meaningful contexts without creating unnecessary barriers to understanding the mathematical problem. In particular, cultural knowledge should not become an unintended prerequisite for answering a numeracy question unless that knowledge is explicitly part of the competency being assessed.

Future research should examine whether items using different cultural contexts function similarly across student groups. Differential Item Functioning (DIF) analysis, where the sample size

P – ISSN : 2715 – 7067
E – ISSN : 2720 – 8850

and data structure permit, may help identify whether students from different cultural backgrounds with comparable numeracy abilities have different probabilities of answering particular items correctly. Such analysis would strengthen the evidence that the instrument measures numeracy rather than differences in cultural familiarity.

The Utility of the Instrument for Elementary School Teachers

The instrument also offers practical implications for elementary school teachers. The needs analysis showed that 82% of participating teachers had never used numeracy assessment instruments explicitly integrating local cultural contexts. This indicates an opportunity to provide teachers with alternative assessment resources that connect curriculum competencies with students' everyday environments.

The developed instrument can serve as a supplementary assessment resource for evaluating students' numeracy competencies through contextual problems. Rather than relying exclusively on abstract numerical exercises, teachers can use familiar cultural representations to introduce mathematical reasoning, interpretation, measurement, and problem-solving activities.

The student-response questionnaire further indicated that 87% of students reported feeling interested in and familiar with the cultural contexts included in the assessment items. This finding suggests that culturally contextualized questions may encourage student engagement. However, student interest and perceived familiarity should be distinguished from actual improvements in numeracy achievement. The present study primarily provides evidence concerning instrument development and quality, rather than demonstrating a causal effect on learning outcomes.

For practical implementation, teachers can use the instrument to identify students' strengths and difficulties across the three numeracy domains. The item-level analysis can also inform teachers about which questions require adjustment to suit students' learning characteristics and cultural environments. Nevertheless, teachers should use only items that have met the relevant validity and reliability requirements, while the three problematic items should remain excluded or clearly identified as requiring further revision until they have been retested.

At the Disseminate stage, the instrument can be introduced to teachers through limited dissemination and professional development activities. These activities may include guidance on administering the assessment, applying the scoring rubrics, interpreting students' results, and adapting cultural contexts without changing the targeted numeracy competencies. Such support is important to ensure that the instrument is used consistently and appropriately in classroom assessment.

Overall, the findings suggest that local culture-based numeracy assessment has potential as a practical resource for elementary education. Its contribution lies not only in providing alternative assessment questions but also in encouraging teachers to recognize local cultural knowledge as a meaningful context for developing and assessing numeracy competencies.

CONCLUSION

This study successfully developed a North Sumatran local culture-based numeracy assessment instrument for elementary school students, consisting of 27 items covering three numeracy domains: Numbers, Geometry and Measurement, and Data and Uncertainty. The

instrument incorporates cultural contexts from the Batak Toba, Karo, and Mandailing communities to connect numeracy competencies with students' sociocultural environments and everyday experiences.

The expert evaluation results demonstrated a high level of content validity, with scores of 90.9% for content, 87.3% for construction, 88.6% for language, and 93.6% for cultural appropriateness. The field-trial results also indicated promising empirical quality, with a composite reliability coefficient of 0.84. Furthermore, 24 of the 27 items (88.9%) met the established empirical validity criteria. The distribution of item difficulty and the mean discrimination index of 0.38 also indicated that the instrument generally demonstrated acceptable assessment characteristics.

Despite these positive findings, the entire instrument cannot yet be considered fully validated or ready for widespread implementation. Three items did not meet the established empirical validity criteria and therefore require further examination, revision, or elimination. Their performance also highlights the importance of ensuring that cultural contexts are sufficiently accessible and understandable to students from different sociocultural backgrounds. Accordingly, the demonstrated empirical quality of the instrument should be interpreted in relation to the 24 items that met the validity criteria, rather than being generalized to all 27 items.

For further development, future research should conduct pilot testing in elementary schools with different cultural and geographical backgrounds to examine the instrument's applicability across diverse student populations. Further studies should also investigate potential item bias, particularly whether differences in cultural familiarity influence students' responses to numeracy questions. In addition, the three problematic items should be revised or eliminated, followed by further empirical testing to reassess their validity, reliability, and overall contribution to instrument quality. These steps are expected to strengthen the instrument's psychometric evidence and support its broader use as a culturally responsive numeracy assessment tool in Indonesian elementary education.

REFERENCES

Ali, J. A., Abdi, M. K., Ali, T. A., Muse, A. H., & Cumar, M. A. (2025). Geographic and school-level disparities as primary predictors of numeracy skills: A supervised machine learning approach of Somaliland's national learning assessment. *Social Sciences and Humanities Open*, *12*. https://doi.org/10.1016/j.ssaho.2025.102305

Arifa, M. F., Wahdah, N., Fitria, T., Fitria, R., & Jumadi, J. (2023). PBL-Based PhET Simulations e-Worksheet: Improvement students' Communication Skill on Global Warming in Merdeka Curriculum. *ACM International Conference Proceeding Series*, 172–177. https://doi.org/10.1145/3631991.3632019

Bamber, P. (2019). Teacher education for sustainable development and global citizenship: Critical perspectives on values, curriculum and assessment. In *Teacher Education for Sustainable Development and Global Citizenship: Critical Perspectives on Values, Curriculum and Assessment*. Taylor and Francis. https://doi.org/10.4324/9780429427053

Barumbun, M., Salim S, S., Bua, M. T., & Ramadhani, T. (2023). *DEVELOPING A NUMERACY DIAGNOSTIC ASSESSMENT MODULE FOR EARLY PRIMARY SCHOOLERS POST COVID-19 PANDEMIC*. *15*(1). https://doi.org/10.14421/al-bidayah.v15i1.831

Bellini, D., Crescentini, A., Zanolla, G., Cubico, S., Favretto, G., Faccincani, L., Ardolino, P., & Gianesini, G. (2019). Mathematical Competence Scale (MCS) for primary school: The

psychometric properties and the validation of an instrument to enhance the sustainability of talents development through the numeracy skills assessment. *Sustainability (Switzerland)*, *11*(9). https://doi.org/10.3390/su11092569

Calvo, M. B., Martínez-Abad, F., & García, M. E. H. (2014). Selection of key standards to create an instrument for information literacy assessment in compulsory secondary education. In F. J. Garcia-Penalvo & R. I. for E. S. (IUCE) University of Salamanca Department of Computer Science and Automatics Plaza de los Caidos s/n Salamanca (Eds.), *ACM International Conference Proceeding Series* (pp. 475–482). Association for Computing Machinery. https://doi.org/10.1145/2669711.2669942

Cheruiyot, L. G., & Molnár, G. (2026). Technology-based assessment of collaborative problem-solving skills: a bibliometric analysis and review. *Journal of Computers in Education*, *13*(1), 389–410. https://doi.org/10.1007/s40692-025-00361-2

Citrawan, I. W., Mukminin, A., Widana, I. W., Sumandya, I. W., Widana, I. N. S., Arief, H., Razak, R. A., Hadiana, D., & Meter, W. (2024). SPECIAL EDUCATION TEACHERS' ABILITY IN LITERACY AND NUMERACY ASSESSMENTS BASED ON LOCAL WISDOM. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, *8*(1), 145–157. https://doi.org/10.22437/jiituj.v8i1.32608

de Bruyn, C., Said, F. B., Venter, M., & Castanho, R. A. (2026). Are smart technologies enough to build climate-resilient cities? A bibliometric assessment of global trends and research gaps. *City and Environment Interactions*, *29*. https://doi.org/10.1016/j.cacint.2026.100306

Dewanti, S. S., Rahmawati, & Sumardi. (2025). Problem-Based Learning with Manipulative Media: Enhancing Elementary Students' Conceptual Understanding in Area and Volume. *Jurnal Ilmiah Sekolah Dasar*, *8*(3), 487–499. https://doi.org/10.23887/jisd.v8i3.80221

Eraikhuemen, L., & Aikpitanyi, L. A. (2015a). Mathematics teachers' awareness of ethnomathematics approach in teaching. *International Journal of Assessment and Evaluation*, *21*(3–4), 1–8. https://doi.org/10.18848/2327-7920/cgp/v21i3-4/48364

Eraikhuemen, L., & Aikpitanyi, L. A. (2015b). Mathematics teachers' awareness of ethnomathematics approach in teaching. *International Journal of Assessment and Evaluation*, *21*(3–4), 1–8. https://doi.org/10.18848/2327-7920/cgp/v21i3-4/48364

Essiet, I. A., Warner, E., Lander, N. J., Salmon, J., Duncan, M. J., Eyre, E. L. J., & Barnett, L. M. (2023). Primary School Teachers' Perceptions of Physical Literacy Assessment: A Mixed-Methods Study. *Journal of Teaching in Physical Education*, *42*(4), 609–620. https://doi.org/10.1123/jtpe.2022-0091

Fu, G., Hu, S., & Qu, Y. (2026). Analysis of Trends in Screening, Assessment, and Monitoring of Gestational Diabetes Mellitus Through Bibliometric Analysis. *Journal of Diabetes Research*, *2026*(1). https://doi.org/10.1155/jdr/6825254

Hortsch, M., Girão-Carmona, V. C. C., Leite, A. C. R. de M., Nikas, I. P., Gatumu, M. K., Koney, N. K.-K., Arko-Boham, B., Yohannan, D. G., Oommen, A. M., Li, Y., Yang, J., Trollope, A. F., Meyer, A. J., & Van Nuland, S. E. (2026). A global overview of anatomical science education and its present and future role in biomedical curricula. *Anatomical Sciences Education*, *19*(1), 5–45. https://doi.org/10.1002/ase.70137

Karatepe Mumcu, Y. (2026). Deployment Readiness of Artificial Neural Networks in Power Systems (2020–2024): A Bibliometric and Engineering Assessment Using a Domain-Level Evaluation Framework. *Energies*, *19*(7). https://doi.org/10.3390/en19071610

Kelana, J. B., Wardani, D. S., Firdaus, A. R., Altaftazani, D. H., & Rahayu, G. D. S. (2020). The effect of STEM approach on the mathematics literacy ability of elementary school teacher education students. In H. Hendriana, W. Hidayat, S. A. Widodo, M. Irfan, M. S. Noto, K.

S. Perbowo, & R. C. I. Prahmana (Eds.), *Journal of Physics: Conference Series* (Vol. 1657, Number 1). IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1657/1/012006

Leite, S. (2024). Weaving Curriculum, Assessment, and Pedagogy: Global Citizenship Experience Lab School's Whole-School Approach to Sustainability and Global Citizenship Education. *Sustainable Development Goals Series*, *Part F3104*, 215–226. https://doi.org/10.1007/978-3-031-56172-6_15

Munadi, S., & Ratna Febriyanti, W. D. (2020). Design and validation of mathematical literacy instruments for assessment for learning in Indonesia. *European Journal of Educational Research*, *9*(2), 865–875. https://doi.org/10.12973/eu-jer.9.2.865

Niklas, F., Birtwistle, E., Wirth, A., Schiele, T., & Mues, A. (2022). App-based learning for kindergarten children at home (Learning4Kids): Study protocol for cohort 2 and the school assessments. *BMC Pediatrics*, *22*(1). https://doi.org/10.1186/s12887-022-03737-w

No, H. J., Choi, H., Hong, S.-G., Kwak, S.-Y., & Cho, S. (2012). Integrating global education program into engineering curriculum: Developing global engineering education program at GECE. *Proceedings of IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2012*, T1D13-T1D17. https://doi.org/10.1109/TALE.2012.6360378

Randy, M. Y., Sukoriyanto, S., Rahayuningsih, S., Dewi, R. S. I., & Kusumaningrum, S. R. (2024). DEVELOPMENT OF A NUMERACY LITERACY BASED MATHEMATICS TEST ASSESSMENT ON INTEGERS IN PRIMARY SCHOOLS. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *13*(4), 1202. https://doi.org/10.24127/ajpm.v13i4.9797

Richards, J., & Islam, M. S. (2018). Assessing literacy and numeracy among primary school students: A pilot survey in rural Bangladesh. *International Journal of Educational Development*, *61*, 55–63. https://doi.org/10.1016/j.ijedudev.2017.12.001

Starmer, D. L., Chapman, E., & Millward, M. J. (2010). Applying global frameworks to assessment in medical education: An example of a nationally produced curriculum for cancer education. *Journal of Cancer Education*, *25*(3), 285–289. https://doi.org/10.1007/s13187-010-0096-8

Su, X., Chau, K. Y., Ho, G. T. S., Yip, H. T., & Tang, Y. M. (2026). A bibliometric study on technology usage for occupational safety and health risk assessment in construction industry. *Journal of Asian Architecture and Building Engineering*, *25*(3), 2400–2415. https://doi.org/10.1080/13467581.2025.2499727

Suherman, S., & Vidákovich, T. (2022). Tapis Patterns in the Context of Ethnomathematics to Assess Students' Creative Thinking in Mathematics: A Rasch Measurement. *Mathematics Teaching-Research Journal*, *14*(4), 56–79. https://www.scopus.com/pages/publications/85144736487?origin=resultslist

Sun, T., & Buripakdi, A. (2024). Cultural Education in Transition: Re-envisaging National English Curricula of China from a Global Citizenship Perspective. *Asia-Pacific Education Researcher*, *33*(6), 1433–1444. https://doi.org/10.1007/s40299-024-00824-7

Urwick, J. (2022). Dimensions of inequality in children's literacy and numeracy in Uganda: Evidence from a household-based assessment. *International Journal of Educational Development*, *88*. https://doi.org/10.1016/j.ijedudev.2021.102525

Wang, Y., Zheng, S., Yang, Z., Guo, J., Yang, Z., & Hong, J. (2026). A global and local agent-based curriculum reinforcement learning approach for multi-end-effector robotic arm manipulation. *Engineering Applications of Artificial Intelligence*, *163*. https://doi.org/10.1016/j.engappai.2025.113121

Zulaeha, O., & Sangadji, H. (2025). Authentic Assessment of Literacy and Numeracy Based on Local Wisdom at Elementary School Level in Ternate City. *Journal Evaluation in Education (JEE)*, *6*(4), 1141–1150. https://doi.org/10.37251/jee.v6i4.2035