



Developing Algebra-Based Mathematical Literacy Assessment Items Using Individual and Cultural Contexts for Grade VIII Students

Alya Fadhila, M. Duskri*

Universitas Islam Negeri Ar-Raniry, Banda aceh, Indonesia

ARTICLE INFO

Article History

Received :30-04-2026

Revised :03-06-2026

Accepted :05-08-2026

Keywords:

*Algebra Content;
Assessment Instrument;
Individual and Acehese Cultural
Contexts;
Mathematical Literacy;
PISA Framework.*

Doi:

<https://doi.org/10.24042/djm.v9i2.31617>

ABSTRACT

Mathematical literacy assessment should capture students' capacity to use mathematics in meaningful situations while maintaining sound measurement quality. This study aimed to develop and empirically evaluate an algebra-based mathematical literacy assessment for Grade VIII students that incorporates individual and Acehese cultural contexts across different items within a single instrument. A Research and Development approach was employed using Tessmer's formative evaluation framework, comprising preliminary analysis, self-evaluation, prototyping, and field testing. The one-to-one evaluation involved 6 students, the small-group evaluation involved 34 students, and the field test involved 128 students from four junior secondary schools in Banda Aceh. The final instrument consisted of 13 multiple-choice items, including 5 individual-context items and 8 Acehese cultural-context items, and was evaluated through expert judgment and classical item analysis using ITEMAN. Expert validation yielded an overall mean score of 4.40, indicating strong content and contextual appropriateness. Field-test results showed that all 13 items were retained, with point-biserial coefficients ranging from 0.299 to 0.609 and an alpha reliability coefficient of 0.712. Item difficulty analysis identified 11 moderate items, 1 easy item, and 1 difficult item, while discrimination analysis classified 10 items as very good, 2 as good, and 1 as fair. Of the 39 distractors, 38 (97.44%) functioned effectively, while student responses indicated good attractiveness (71.1%) and usefulness (79%) and moderate perceived difficulty (55.6%). These findings indicate that individual and Acehese cultural contexts can be systematically represented across an algebra-based mathematical literacy assessment while maintaining acceptable psychometric characteristics and favorable responses from the participating students.

<http://ejournal.radenintan.ac.id/index.php/desimal/index>

Correspondence author

M. Duskri, Universitas Islam Negeri Ar-Raniry, Banda aceh, Indonesia
email: m.duskri@ar-raniry.ac.id

INTRODUCTION

Mathematical literacy has become an essential educational competence because students are increasingly expected to use mathematics not only within the classroom but also to interpret information, reason about quantitative situations, and make informed decisions in everyday life. Within the Programme for International Student Assessment (PISA), mathematical literacy emphasizes the capacity to formulate, employ, and interpret mathematics across situations that extend beyond routine procedural exercises (OECD, 2023). This orientation positions mathematics as a tool for understanding and responding to real-world problems rather than merely a collection of formulas and computational procedures. Recent research has consequently placed greater emphasis on assessments that require students to connect mathematical knowledge with meaningful situations, interpret contextual information, and determine appropriate mathematical strategies (Çakıroğlu et al., 2024; Koskinen & Pitkaniemi, 2022). Such competencies are particularly important at the junior secondary level, when students begin to encounter mathematical ideas that demand greater abstraction while remaining applicable to familiar situations. Therefore, the quality of mathematical literacy assessment depends not only on the mathematical content being tested but also on whether the items provide contexts through which students can meaningfully demonstrate their reasoning.

The need for appropriate mathematical literacy assessment is particularly relevant in Indonesia, where students continue to experience difficulties in solving contextual and reasoning-oriented mathematics problems. PISA 2022 results indicate that Indonesian students' mathematics performance remains below the OECD average, reflecting persistent challenges in applying mathematical knowledge to unfamiliar and real-world situations (OECD, 2023). Similar concerns have been identified in studies of Indonesian students' numeracy and mathematical literacy, particularly when problems require interpretation, modeling, and reasoning beyond direct computation (Pratiwi et al., 2024). These difficulties suggest that conventional items emphasizing procedural execution may provide an incomplete picture of what students can actually do with mathematics. At the same time, assessment instruments need to distinguish between difficulties arising from mathematical concepts and those caused by students' unfamiliarity with the situation in which a problem is presented. Developing assessment items that combine sound psychometric characteristics with meaningful contexts is therefore necessary to obtain a more informative representation of students' mathematical literacy.

Among the mathematical domains encountered in junior secondary education, algebra has a particularly important role in the development of mathematical literacy. Algebra enables students to represent relationships among quantities, recognize patterns, formulate generalizations, and translate real situations into mathematical models. These activities are closely connected with the PISA processes of formulating situations mathematically, employing mathematical concepts and procedures, and interpreting mathematical results in relation to the original problem (OECD, 2023). For Grade VIII students, algebra also represents an important transition from predominantly arithmetic reasoning toward more abstract forms of mathematical thinking involving variables, expressions, equations, and relationships. Recent instrument-development studies have

therefore identified algebra as an important domain for measuring students' capacity to move between contextual information and symbolic mathematical representations (Hwang, 2025; Sorby et al., 2022). However, assessing algebraic literacy requires more than placing algebraic operations inside word problems; the context itself must provide a meaningful reason for students to formulate, use, and interpret algebra. This creates a need for algebra assessment items that preserve mathematical rigor while connecting abstract relationships with situations that students can recognize and interpret.

The relevance of an assessment context becomes especially important when mathematical literacy is viewed as the application of mathematics across different dimensions of students' lives. Individual contexts can situate algebraic reasoning in experiences such as purchasing goods, managing pocket money, planning activities, or making personal financial decisions, thereby providing situations in which quantities and relationships have recognizable meanings (Ammer & Aldhyani, 2022; Shah & Shah, 2023). Cultural contexts offer another dimension by embedding mathematical relationships in local practices, products, traditions, and forms of knowledge that are familiar within a particular community. Culturally responsive perspectives in mathematics education emphasize that students' cultural experiences should not be treated as separate from academic mathematics but can provide meaningful resources for teaching and assessment (Kolovou, 2023; Nolan & Xenofontos, 2023). In Aceh, cultural elements such as traditional foods, local commodities, and community traditions provide potential settings for constructing mathematical problems without changing the underlying algebraic competencies being assessed. Recent studies also indicate growing interest in connecting mathematical tasks with local and ethnomathematical contexts because such tasks can make school mathematics more closely related to students' social environments (Asare, 2026; Kabuye Batiibwe, 2024). Accordingly, incorporating both personal experience and local culture at the instrument level may broaden the contextual range through which students are invited to demonstrate mathematical literacy.

Previous research has established a substantial foundation for the development of contextual mathematical literacy assessments, although the studies vary considerably in content, population, and contextual orientation. Nurgabyl et al., (2023) developed mathematical literacy items based on the PISA framework using contextual situations, while Fitriza et al., (2024) demonstrated the potential of Minangkabau cultural elements in mathematical literacy problems. Studies of numeracy and contextual problem solving have similarly emphasized the importance of presenting mathematics through situations that require interpretation and reasoning rather than direct procedural recall (Choi et al., 2026; Nilimaa, 2023). From an assessment perspective, Chen et al., (2023); D'Erchie et al., (2026) examined mathematical literacy instruments in algebra using item response theory, illustrating the importance of establishing defensible item characteristics when literacy is measured through test instruments. More recently, Utami, (2025) developed algebra-based test items for measuring mathematical literacy and reported evidence of validity and reliability, confirming that algebra can be systematically operationalized within a literacy-oriented assessment. Research on culturally responsive mathematics has meanwhile reinforced the value of connecting formal mathematics with students' sociocultural

experiences Banjo et al., (2025), while contextual and culturally grounded studies have continued to demonstrate the educational relevance of locally meaningful mathematical situations (Fouze & Amit, 2025; Legarde, 2026). Collectively, these studies show two dominant directions in the literature: psychometrically oriented development of mathematical literacy instruments and contextually oriented development of tasks grounded in everyday or cultural experiences.

Despite these developments, an important gap remains at the intersection of algebra content, contextual diversity, and assessment instrument development. Existing algebra-focused instruments have primarily concentrated on the measurement properties of mathematical literacy items without making the combination of students' individual experiences and specific local cultural settings a central feature of the instrument (Dickson, 2025). Conversely, culturally oriented studies have generally concentrated on a particular cultural setting, such as Minangkabau or Acehnese culture, rather than deliberately placing cultural items alongside individual-context items within the same algebra assessment (Choi et al., 2026; Nilimaa, 2023). This distinction is especially important because recent work has already demonstrated the feasibility of developing mathematical literacy problems based on Acehnese ethnomathematics; consequently, the contribution of the present study cannot be defined merely by the use of Acehnese culture. What remains less explored is an instrument-level design in which Grade VIII students encounter both individual contexts and Acehnese cultural contexts across different items while the items consistently target algebraic mathematical literacy. Moreover, such an instrument requires empirical evidence concerning item validity, reliability, difficulty, discrimination, and distractor functioning before it can be considered suitable for assessment purposes. Addressing these limitations provides a more precise basis for positioning contextual diversity as part of assessment design rather than treating context only as a pedagogical illustration.

Against this background, this study aims to develop and evaluate an algebra-based mathematical literacy assessment instrument for Grade VIII students that incorporates individual contexts and Acehnese cultural contexts across different items within a single instrument. The study specifically examines the quality of the developed items through expert evaluation and empirical analysis of item validity, test reliability, difficulty indices, discrimination indices, and distractor effectiveness. Theoretically, the study contributes to mathematical literacy assessment by demonstrating how different contextual categories can be systematically represented within one algebra assessment while remaining aligned with mathematical literacy processes. It also extends culturally responsive perspectives from the use of local culture alone toward an assessment design that places cultural situations alongside students' personal and everyday experiences. Practically, the resulting instrument is intended to provide mathematics teachers with an alternative tool for assessing students' ability to formulate, employ, and interpret algebra in situations that are connected to both their individual lives and cultural environment. The instrument may also provide a development model for researchers seeking to construct contextual mathematical literacy assessments using other mathematical domains and regional cultures. In this way, the study seeks to contribute not simply another set of contextual

mathematics problems, but a psychometrically evaluated assessment that broadens the contexts through which students' algebraic mathematical literacy can be examined.

METHOD

Research Design

This study employed a Research and Development (R&D) approach to develop and empirically evaluate a mathematical literacy assessment instrument on algebra content incorporating individual and Acehnese cultural contexts for Grade VIII students. The study was designed as a development study using Tessmer's formative evaluation framework, which emphasizes iterative product refinement through successive evaluation involving experts and intended users (Elmalı & Çıtak, 2026; Schexnayder et al., 2023). This framework was considered appropriate because the development of assessment instruments requires systematic examination of content appropriateness, readability, contextual relevance, and empirical item functioning before the instrument is finalized. The development process comprised four main stages: preliminary analysis, self-evaluation, prototyping, and field testing, with the prototyping stage incorporating expert review, one-to-one evaluation, and small-group evaluation. Evidence obtained at each stage was used to evaluate and, where necessary, revise the instrument before proceeding to the subsequent stage. Accordingly, the development process was intended to establish both the content appropriateness and empirical quality of the resulting mathematical literacy assessment.

Research Participants and Setting

The study was conducted during the 2026/2027 academic year at MTsN 1 Banda Aceh, SMP Negeri 1 Banda Aceh, SMP Negeri 3 Banda Aceh, and SMP Negeri 17 Banda Aceh. The schools were selected purposively to provide variation in school characteristics, including an Islamic junior secondary school, a high-achieving school, and schools serving students from diverse socioeconomic backgrounds. Participants were Grade VIII students who had studied the algebra content required to respond to the developed assessment items. At the one-to-one stage, six students were purposively selected based on mathematics teachers' recommendations to represent high, medium, and low academic ability, whereas the small-group evaluation involved 34 students from one class. The field test involved 128 Grade VIII students from the four participating schools and provided the principal response data for empirical item analysis. The progressive involvement of students from smaller to larger groups was consistent with the formative evaluation principle of examining a developing product through successive evaluation stages before its final empirical assessment (Kaya & Kaya, 2024; Monteiro & Passarinho, 2026).

Development Procedure

The development procedure consisted of four interconnected stages: preliminary analysis, self-evaluation, prototyping, and field testing. During the preliminary stage, a literature review was conducted on mathematical literacy, the PISA mathematics framework, Grade VIII algebra content, and individual and Acehnese cultural contexts. Curriculum analysis was undertaken by examining the learning outcomes for algebra in

Phase D of the Merdeka Curriculum to establish the curricular scope of the assessment. Context analysis was also conducted to identify situations from students' individual experiences and the Acehese cultural environment that could meaningfully accommodate algebraic problems. Individual contexts included situations related to students' personal experiences, whereas cultural contexts incorporated locally relevant elements such as traditional foods, local commodities, and Acehese social traditions. The results of these analyses provided the conceptual, curricular, and contextual basis for constructing the test specifications and the initial item prototype.

During self-evaluation, the researchers developed item indicators, test specifications, answer alternatives, scoring keys, and the initial set of assessment items. The mathematical literacy dimension was guided by the PISA mathematics framework, particularly the processes of formulating situations mathematically, employing mathematical concepts, facts, and procedures, and interpreting and evaluating mathematical outcomes (OECD, 2023). The 13 items were also distributed across cognitive demand levels in the Revised Bloom's Taxonomy, consisting of one C2 item (7.6%), four C3 items (30.8%), six C4 items (46.2%), and two C5 items (15.4%). The Revised Bloom's Taxonomy and PISA mathematical literacy processes were treated as complementary rather than equivalent frameworks: Bloom's taxonomy characterized the cognitive demand of the items, whereas the PISA framework specified the mathematical literacy processes elicited by the contextual problems. Consequently, no direct one-to-one correspondence was assumed between a particular Bloom cognitive level and an individual PISA mathematical process. Both dimensions were incorporated into the test specification to provide variation in cognitive demand while maintaining alignment with the mathematical literacy construct.

The prototyping stage began with an expert review of Prototype I by three validators with relevant professional expertise. The panel consisted of one university lecturer specializing in mathematics education, assessment development, and mathematical literacy and two junior secondary mathematics teachers experienced in algebra instruction and student assessment. The validators examined content appropriateness, construction, language clarity, mathematical literacy alignment, item difficulty, visual presentation, and the relevance and accuracy of the individual and Acehese cultural contexts. Their quantitative judgments and qualitative comments were used to identify content, linguistic, contextual, and presentation issues that required modification before the instrument was administered to students. The revised prototype was then administered to six students from class VIII-3 of SMP Negeri 1 Banda Aceh during the one-to-one evaluation to examine readability, clarity of instructions, understanding of contextual information, and difficulties experienced while answering the items. Feedback obtained during this stage informed further refinement of the instrument before the small-group evaluation.

The small-group evaluation involved 34 students from class VIII-11 of MTsN 1 Banda Aceh who had studied the relevant algebra content. This stage provided an initial examination of item functioning under classroom conditions and was also used to identify remaining issues concerning wording, contextual interpretation, answer alternatives, and item difficulty. Following the small-group evaluation and necessary revisions, the field test was conducted with 128 Grade VIII students from four classes: VIII-10 of MTsN 1 Banda

Aceh, VIII-1 of SMP Negeri 1 Banda Aceh, VIII-1 of SMP Negeri 3 Banda Aceh, and VIII-7 of SMP Negeri 17 Banda Aceh. Field-test responses constituted the primary empirical dataset used to examine the characteristics of the developed assessment through classical item analysis. The analysis focused on item-total relationships, internal consistency reliability, difficulty indices, discrimination indices, and distractor functioning. Thus, the field test served as the final empirical evaluation of the instrument within the participating sample before conclusions regarding its psychometric quality were drawn.

Research Instruments

Four instruments were used throughout the development process: an expert validation sheet, the mathematical literacy test, an unstructured interview guide, and a student response questionnaire. The expert validation sheet was designed to assess the content, construction, language, mathematical literacy alignment, and appropriateness of the individual and Acehnese cultural contexts using a five-point rating scale. Each validator independently evaluated the prototype and provided both numerical ratings and qualitative comments to support item revision. The five response categories ranged from 1, indicating the lowest level of agreement, to 5, indicating the highest level of agreement, as presented in Table 1. These ratings provided expert-judgment evidence concerning the appropriateness of the instrument before student testing. The qualitative comments were considered alongside the quantitative ratings when determining which aspects of Prototype I required revision.

Table 1. Validator Assessment Categories

Score	Assessment Category
1	Strongly disagree
2	Disagree
3	Fairly good
4	Agree
5	Strongly agree

The principal product of the study was a 13-item multiple-choice mathematical literacy test covering Grade VIII algebra and incorporating individual and Acehnese cultural contexts across different items within a single assessment instrument. Five items were developed using individual contexts and eight using Acehnese cultural contexts, with each item employing only one type of context rather than combining both contexts within the same problem. Each multiple-choice item consisted of one correct answer and three distractors, resulting in 39 distractors across the complete assessment. The unstructured interview guide was used during formative evaluation to obtain information concerning students' interpretation of instructions and contexts, readability of the items, and difficulties encountered while solving the problems. In addition, the student response questionnaire used a five-point Likert scale to collect supplementary information concerning the readability, attractiveness, perceived difficulty, and usefulness of the developed assessment. The questionnaire was administered to 23 purposively selected students after the field test and served as supporting user-oriented evidence rather than the primary basis for evaluating psychometric quality.

Data Collection

Data collection followed the successive stages of formative evaluation so that evidence obtained at one stage could inform evaluation and refinement at the next stage. During expert review, the three validators independently assessed Prototype I and provided quantitative ratings and qualitative feedback concerning content, construction, language, mathematical literacy alignment, and contextual appropriateness. During the one-to-one and small-group stages, students completed the assessment and provided feedback through unstructured interviews regarding item readability, clarity of instructions, contextual understanding, and difficulties encountered during problem solving. At the field-test stage, the revised instrument was administered to 128 Grade VIII students, and their item responses were recorded for empirical analysis of item characteristics. Following the field test, the student response questionnaire was administered to 23 purposively selected students to obtain supplementary evidence concerning the readability, attractiveness, perceived difficulty, and usefulness of the instrument. Questionnaire responses were interpreted as supporting evidence regarding user acceptability and practicality and were not used to determine the psychometric quality of the assessment.

Data Analysis

Data analysis combined expert-judgment evidence with empirical analysis of student responses. Expert validation ratings were summarized descriptively across the evaluated dimensions, while the validators' written comments were examined to identify aspects requiring revision. To obtain a standardized validation score, the total score obtained from the validators was compared with the maximum possible score and expressed as a percentage. The resulting percentage was subsequently interpreted according to the validation categories presented in Table 2. This procedure enabled quantitative expert ratings to be interpreted consistently while retaining qualitative comments as an additional basis for item refinement. The expert-validation evidence was therefore used primarily to establish the appropriateness of the instrument before empirical testing rather than as a substitute for item-level statistical analysis.

Table 2. Validation Score Categories

Score Interval	Category
0–20	Very invalid
21–40	Invalid
41–60	Less valid
61–80	Valid
81–100	Very valid

Field-test responses were analyzed using ITEMAN to examine the empirical functioning of the 13 multiple-choice items within a classical test analysis framework. Because the items were dichotomously scored as correct or incorrect, point-biserial coefficients (rpbis) were used to examine the relationship between performance on individual items and performance on the test as a whole. This approach was maintained consistently throughout the empirical item analysis rather than combining point-biserial and product-moment criteria. Test-level internal consistency was evaluated using the

alpha reliability coefficient, and $\alpha \geq .70$ was treated as the minimum criterion for acceptable internal consistency. Reliability coefficients were additionally interpreted descriptively according to the categories shown in Table 3. Thus, the decision regarding test reliability was distinguished from the descriptive classification of the magnitude of the coefficient.

Table 3. Reliability Level Categories

Alpha Coefficient (α)	Category
$0.00 \leq \alpha < 0.20$	Very low reliability
$0.20 \leq \alpha < 0.40$	Low reliability
$0.40 \leq \alpha < 0.60$	Moderate reliability
$0.60 \leq \alpha < 0.80$	High reliability
$0.80 \leq \alpha \leq 1.00$	Very high reliability

Item difficulty was determined from the proportion of field-test participants who answered each item correctly, expressed as the difficulty index (p). Higher p values indicate easier items, whereas lower values indicate more difficult items. Items with difficulty indices from .30 to .70 were regarded as moderate and fell within the preferred range, while items outside this interval were examined together with their discrimination and distractor statistics before a retention or revision decision was made (Rezigalla et al., 2024). This approach avoided evaluating an item solely on the basis of its difficulty because relatively easy or difficult items may still provide useful measurement information when they demonstrate adequate discrimination and content relevance. The classification and decision criteria applied to the difficulty indices are presented in Table 4. These criteria were subsequently used consistently when interpreting item-level results from the small-group evaluation and field test.

Table 4. Item Difficulty Index Categories

Difficulty Index (p)	Category	Decision
0.91–1.00	Very Easy	Rejected
0.71–0.90	Easy	Accepted/Revised
0.30–0.70	Moderate	Accepted
0.10–0.29	Difficult	Accepted/Revised
0.00–0.09	Very Difficult	Rejected

Item discrimination was examined to determine the extent to which each item differentiated between students with higher and lower overall test performance. Point-biserial coefficients obtained from the ITEMAN analysis were used as the principal indicator because the items were dichotomously scored while the total test score represented a continuous measure of overall performance. Coefficients of .30 or higher were regarded as acceptable, whereas values from .20 to .29 indicated that an item required further examination or revision. Substantially lower or negative coefficients were interpreted as evidence of problematic item functioning and required careful examination of item content, answer keys, wording, and distractors. Decisions concerning retention or revision were therefore based on the discrimination coefficient together with difficulty, content considerations, and distractor performance rather than on a single statistic in

isolation. The criteria used to interpret the discrimination indices are summarized in Table 5.

Table 5. Classification of Discrimination Index Interpretation

Discrimination Index	Category	Decision
0.40–1.00	Very good	Accepted
0.30–0.39	Good	Accepted without revision
0.20–0.29	Fair	Revision
–1.00–0.19	Very poor	Rejected

Distractor effectiveness was evaluated for each incorrect answer option of the multiple-choice items. A distractor was classified as functioning when it was selected by at least 5% of the field-test participants, whereas an option selected by fewer than 5% was examined as a potentially nonfunctioning distractor. Because each of the 13 items contained three incorrect alternatives, the distractor analysis covered a total of 39 options. Distractor functioning was interpreted together with difficulty and discrimination because plausible incorrect alternatives should attract students who have not yet mastered the concepts or reasoning processes required by an item. The ITEMAN output was used to summarize the relevant item statistics and provide empirical evidence for decisions concerning item retention and revision. Consequently, the final judgment of instrument quality was based on converging evidence from expert validation, internal consistency, item difficulty, item discrimination, and distractor effectiveness.

Ethical Considerations

The study involved Grade VIII students and therefore required attention to voluntary participation, confidentiality, and the protection of student information throughout data collection and reporting. Research activities were conducted with permission from the participating schools and in accordance with the procedures applicable at the participating institutions. Students were informed about the purpose and procedures of the research, and participation in the study did not affect their academic grades or regular school assessment. Participant identities were kept confidential, and no personally identifying information was included in the analysis or publication of the expert-review, interview, questionnaire, or field-test data. Because the participants were minors, consent and assent procedures should be reported in accordance with the procedures actually implemented by the researchers and the requirements of the participating institutions. All collected data were used solely for research and assessment-development purposes.

RESULTS AND DISCUSSION

Results

Preliminary and Self-Evaluation Stages

The preliminary stage of this study resulted in a product consisting of 13 mathematical literacy items on algebra content with individual and cultural contexts for eighth-grade junior secondary school students. During the preliminary stage, a needs analysis and curriculum review were conducted to ensure that the designed instrument was aligned with students' needs and the demands of basic competencies. The curriculum

analysis was carried out by examining the Learning Outcomes (CP) of Phase D mathematics for eighth-grade junior secondary school students, with a focus on algebra content. The results of the analysis indicated that algebra is an essential content area that students need to master, including the ability to model real-life situations into algebraic forms, solve linear equations, and interpret solutions within the context of real-world problems.

Student analysis was conducted to identify the characteristics of eighth-grade junior secondary school students as prospective users of the product. The research participants consisted of eighth-grade students from several schools in Banda Aceh, with an average age of 13–14 years. According to Jean Piaget's theory of cognitive development, students at this age are in the early formal operational stage (Piaget, 1972). At this stage, students begin to develop abstract and logical reasoning abilities but still require relevant contexts to facilitate understanding. Therefore, the developed items needed to use clear language, present contexts that are closely related to students' lives, include supporting illustrations, and provide a variety of difficulty levels.

Context analysis was conducted to determine the types of contexts to be integrated into the mathematical literacy items. Based on the literature review and the PISA framework, two types of contexts were selected: individual contexts related to students' personal experiences, such as pocket money, hobbies, buying and selling, and personal financial planning, and cultural contexts related to Acehnese local wisdom, such as traditional foods (timphan), local tourism (Gayo coffee), and the meugang tradition. In the developed instrument, each item used only one type of context, either an individual context or an Acehnese cultural context, not both within a single item. Of the 13 developed items, 5 used individual contexts and 8 used Acehnese cultural contexts. The integration of both contexts occurred at the instrument level, where both types of contexts were presented simultaneously within a single assessment instrument. This was intended to train students to apply algebraic concepts in real-life situations while fostering appreciation for Acehnese culture.

During the self-evaluation stage, based on the results of the analyses, Prototype I was developed consisting of 13 multiple-choice items. Each item was developed by the researcher with careful consideration of its alignment with competency achievement indicators, the proportional use of individual and cultural contexts, varying levels of difficulty, the inclusion of supporting illustrations or images, and the clarity of language and instructions. Prototype I was subsequently validated by experts, as described in the following section.

Expert Review

During the expert review stage, Prototype I was validated by three experts. Validator 1 was a university lecturer specializing in mathematics education assessment, Validator 2 was a teacher with expertise in mathematics teaching and assessment practices at the junior secondary school level, and Validator 3 was a teacher with expertise in understanding the characteristics of madrasah students and mathematics instruction at the MTs/SMP level. Each validator assessed the alignment of the content with algebra competencies, alignment with mathematical literacy indicators, relevance of the individual contexts, relevance of the cultural contexts, language clarity, item difficulty level,

appropriateness of visuals or illustrations, item originality, consistency with Tessmer's development model, and suitability for assessment purposes. The evaluation used a scale from 1 to 5, where a score of 1 indicated very inappropriate and a score of 5 indicated very appropriate. A summary of the validation results is presented in Table 6.

Tabel 6. Summary of Expert Validation Results

No.	Assessment Aspect	V1	V2	V3	Average Score
1	Content Appropriateness	4	5	5	4.67
2	Mathematical Literacy Appropriateness	5	5	4	4.67
3	Relevance of Individual Context	5	4	5	4.67
4	Relevance of Cultural Context	5	5	5	5.00
5	Language Clarity	4	3	4	3.67
6	Item Difficulty Level	5	4	5	4.67
7	Visual/Image Appropriateness	3	4	4	3.67
8	Item Authenticity (Originality)	5	3	5	4.33
9	Consistency with the Tessmer Model	5	3	5	4.33
10	Suitability for Assessment	4	4	5	4.33
Overall Average					4.40

Based on Table 6, the instrument was classified as very valid, with an overall average score of 4.40. The aspect with the highest score was the relevance of the cultural context (5.00), indicating that the integration of Acehese cultural elements such as timphan, Gayo coffee, and the meugang tradition was highly authentic and meaningful. The aspects with the lowest scores were language clarity (3.67) and visual appropriateness (3.67), which were still classified as good but required minor improvements, such as simplifying sentence structures and adding supporting illustrations. The validation results and recommendations from the experts were used as the basis for revising Prototype I into Prototype II before it was tested in the subsequent stage. The validators' comments and suggestions regarding one of the developed items are presented in Table 7.

Tabel 7. Revisio Process of Individual Context Items Number 3 and 5

Before Revision
<i>Narasi berikut untuk menjawab soal nomor 3- 5</i>
Jual Beli
Di sebuah pasar tradisional yang penuh warna dan hiruk-pikuk, berbagai pedagang menawarkan aneka buah segar dengan harga yang berbeda-beda. Jeruk segar dijual dengan harga Rp5.000 per kilogram, apel manis dibanderol Rp7.000 per kilogram, dan pisang yang matang ranum tersedia dengan harga Rp4.000 per kilogram. Salah satu pedagang apel sedang memberikan promosi khusus, yaitu diskon 10% untuk setiap pembelian apel lebih dari 3 kilogram. Andi sedang merencanakan untuk membeli beberapa jenis buah untuk keluarganya dengan anggaran yang terbatas. Ia ingin memastikan mendapatkan harga buah yang paling murah sekaligus menghitung total biaya belanjanya agar sesuai dengan dana yang dimiliki.

Dengan memperhatikan harga dan jumlah buah yang tersedia, Andi harus cermat memilih jenis dan banyaknya buah agar pembelian buahnya hemat dan memuaskan semua anggota keluarga. 3. Jika Andi membeli jeruk, apel, dan pisang, fungsi aljabar yang tepat untuk menghitung total biaya adalah...

- A. $5.000x + 7.000y + 4.000z$
- B. $5.000x - 7.000y + 4.000z$
- C. $7.000x + 5.000y + 4.000z$
- D. $5.000 + 7.000 + 4.000 = x + y + z$

5. Berdasarkan harga dan promosi di pasar tersebut, manakah pernyataan berikut yang benar...

- A. Membeli 3 kg apel lebih murah dari 4 kg jeruk
- B. Diskon 10% hanya berlaku jika membeli lebih dari 3 kg apel
- C. Pisang adalah buah paling mahal dalam daftar harga
- D. Harga jeruk 1 kg setara dengan harga pisang 2 kg

Validator Comments

Based on the validation results, the first validator suggested that the instrument be supplemented with relevant illustrations to improve the clarity and attractiveness of the items. Meanwhile, the second and third validators provided several suggestions related to item construction. For item number 3, the validators noted that answer choices A and C did not yet contain statements that clearly distinguished the variables (x), (y), and (z), which could potentially create ambiguity in the interpretation of the answers. In addition, for item number 5, the validators suggested that the item format be changed to a multiple-response multiple-choice format to better align with the measurement objectives and to represent the targeted abilities more comprehensively.

After Revision

Narasi berikut untuk menjawab soal nomor 3- 5

Jual Beli



Sumber: piqsels.com

Di sebuah pasar tradisional, pedagang menawarkan aneka buah segar dengan harga yang berbeda-beda. Jeruk segar dijual dengan harga Rp5.000 per kilogram, apel manis dibanderol Rp7.000 per kilogram, dan pisang yang matang ranum tersedia dengan harga Rp4.000 per kilogram. Salah satu pedagang apel sedang memberikan promosi khusus, yaitu diskon 10% untuk setiap pembelian apel lebih dari 3 kilogram. Andi sedang merencanakan untuk membeli beberapa jenis buah untuk keluarganya dengan anggaran yang terbatas. Ia ingin memastikan mendapatkan harga buah yang paling murah sekaligus menghitung total biaya belanjanya agar sesuai dengan dana yang dimiliki. Dengan memperhatikan harga dan jumlah

buah yang tersedia, Andi harus cermat memilih jenis dan banyaknya buah agar pembelian buahnya hemat dan memuaskan semua anggota keluarga.

3. Jika Andi membeli jeruk, apel, dan pisang, fungsi aljabar yang tepat untuk menghitung total biaya dalam bentuk model aljabar umum adalah....

- A. $(7.000x \times 10\%) + 5.000y + 4.000z$
- B. $7.000x(1 - 10\%) + 5.000y + 4.000z$
- C. $5.000 + 7.000 + 4.000 = x + y + z$
- D. $5.000x + 7.000y + 4.000z$

5. Perhatikan pernyataan-pernyataan berikut:

- I. Kombinasi 5 kg jeruk dan 2 kg pisang selalu lebih murah dibandingkan kombinasi 4 kg apel dan 3 kg pisang.
- II. Diskon 10% hanya didapatkan jika Andi membeli apel dengan berat minimal 3,1 kg atau lebih.
- III. Harga 4 kg pisang sama dengan harga 3,1 kg jeruk.
- IV. Jika Andi membeli 4 kg apel, total harga yang harus dibayar setelah diskon adalah Rp25.200.

Pernyataan-pernyataan diatas yang benar adalah....

- A. I, II, dan III
 - B. I dan III saja
 - C. II dan IV saja
 - D. Semua pernyataan benar
-

All feedback from the validators was used as the basis for revising the instrument, resulting in improved item quality in terms of content, construction, and presentation before it was administered to students.

One-to-One Stage

The one-to-one trial was conducted at SMP 1 Banda Aceh and involved 6 students from class VIII-3 who represented a range of abilities from low to high based on recommendations from the mathematics teacher. The purpose of this stage was to identify the clarity of the language and instructions, the readability of the items and illustrations, the time required to complete all items, and the difficulties experienced by students.

The results of the one-to-one analysis showed that, in general, the items could be read and understood well. Students were able to understand the meaning of the questions and the contexts used. However, several distractors were identified as not functioning optimally because they were too easy to eliminate and did not sufficiently represent possible student errors or misconceptions. Therefore, the distractors were revised to make them more logical, homogeneous, and capable of attracting students who had not yet mastered the concepts. The instructions were generally clear, but they needed to be made more explicit to avoid ambiguity. Overall, the results of this stage indicated that the instrument required minor revisions to improve distractor quality before proceeding to the next stage.

Small Group Stage

The small group trial was conducted at MTsN 1 Banda Aceh and involved 34 students from class VIII-1. This stage aimed to obtain an initial overview of the quality of the items before large-scale field testing. The analysis of the 13 items showed that all items (100%)

were declared valid because the obtained r_{pbis} values ranged from 0.359 to 0.650, exceeding the critical r -value of 0.339. This finding indicates that each item made a significant contribution to the mathematical literacy construct being measured. The increase in the sample size from 6 to 34 students resulted in more stable and significant correlations.

Test reliability was calculated using Cronbach's Alpha, yielding a coefficient of 0.744, which falls within the high category and indicates that the instrument has good internal consistency. The item difficulty levels varied, with 1 item (7.7%) classified as easy (item 1), 11 items (84.6%) classified as moderate (items 2–11 and 13), and 1 item (7.7%) classified as difficult (item 12). This composition was considered ideal for an achievement test because it was dominated by items with a moderate level of difficulty. The discrimination analysis showed that 10 items (76.9%) had good discrimination indices and 3 items (23.1%) had fair discrimination indices, with no items exhibiting poor discrimination. This indicates that all items were capable of distinguishing between students with low and high levels of ability. Therefore, the instrument was considered suitable to proceed to the next stage.

Field Test Stage

The field test was conducted in four schools, namely MTsN 1 Banda Aceh class VIII-10, SMP Negeri 1 Banda Aceh class VIII-1, SMP Negeri 3 Banda Aceh class VIII-1, and SMP Negeri 17 Banda Aceh class VIII-7, involving 128 eighth-grade students. This stage aimed to obtain empirical data regarding the quality of the test items on a broader scale and to evaluate the difficulty index, discrimination index, distractor effectiveness, validity, and reliability more accurately. The results of the analysis are presented in Table 8.

Tabel 8. Summary of the Field Test Analysis

Item No.	Difficulty Index	Category	Biserial	Discrimination Index	Category	Conclusion
1	0.703	Easy	0.409	0.309	Good	Accepted
2	0.336	Moderate	0.592	0.457	Very good	Accepted
3	0.422	Moderate	0.616	0.488	Very good	Accepted
4	0.664	Moderate	0.788	0.609	Very good	Accepted
5	0.516	Moderate	0.688	0.549	Very good	Accepted
6	0.430	Moderate	0.634	0.503	Very good	Accepted
7	0.328	Moderate	0.648	0.499	Very good	Accepted
8	0.416	Moderate	0.451	0.359	Good	Accepted
9	0.430	Moderate	0.628	0.498	Very good	Accepted
10	0.391	Moderate	0.599	0.471	Very good	Accepted
11	0.516	Moderate	0.635	0.507	Very good	Accepted
12	0.273	Difficult	0.401	0.299	Fair	Accepted
13	0.477	Moderate	0.740	0.590	Very good	Accepted

Based on the analysis using the ITEMAN 3.0 program involving 128 students, as presented in Table 8, all 13 items were accepted because they had positive point-biserial values and exceeded the minimum threshold of 0.20. The point-biserial values ranged from 0.299 to 0.609. These results indicate that each item was capable of distinguishing between

students with high and low levels of mathematical literacy and contributed to the measurement of the intended construct.

Although all items met the validity criteria, the quality of the items varied. Item number 4 demonstrated the highest quality, with the highest point-biserial value of 0.609 and a moderate difficulty level ($P = 0.664$). This value indicates that the item has excellent discrimination ability in distinguishing between high- and low-ability students. In addition, the moderate difficulty level indicates that the item can be answered correctly by a substantial proportion of students without losing its ability to identify differences in ability. Therefore, item number 4 was retained without revision and is recommended as one of the most representative items for measuring mathematical literacy in algebra content.

Item number 13 also demonstrated very good quality, with a point-biserial value of 0.590 and a moderate difficulty level ($P = 0.477$). These results indicate that the item was able to measure students' abilities optimally and was therefore retained without revision.

In contrast, item number 12 had the lowest point-biserial value of 0.299 and the highest difficulty level ($P = 0.273$). These results indicate that the item was relatively more difficult than the other items and could be answered correctly by only about 27.3% of the students. The low proportion of correct responses is presumed to be due to the more complex reasoning required to connect contextual information with the appropriate algebraic model. Nevertheless, the point-biserial value remained above the minimum threshold, indicating that the item was still capable of distinguishing students based on their ability levels. Therefore, item number 12 was retained because it plays a role in identifying students with higher levels of mathematical literacy.

Item number 1 had the highest difficulty index ($P = 0.703$) and was therefore classified as easy. Its point-biserial value of 0.309 indicates that the item still possessed adequate discrimination ability, although not as strong as that of most other items. The high proportion of students who answered correctly suggests that the context and solution procedure were relatively easy to understand. Therefore, item number 1 was retained without revision because it functions as an introductory item that can increase students' confidence at the beginning of the test, as presented in Figure 2.

Bacalah narasi berikut untuk menjawab soal nomor 1- 2

Timphan



Sumber: tradisikuliner.com

Timphan merupakan salah satu makanan tradisional khas Aceh yang tidak hanya dikenal karena cita rasanya, tetapi juga karena nilai budaya yang melekat di dalamnya. Dalam kehidupan masyarakat Aceh, timphan sering disajikan pada berbagai momen penting seperti hari raya, kenduri, dan upacara adat sebagai simbol kebersamaan, penghormatan kepada tamu, serta pelestarian tradisi leluhur.

Timphan dibuat dari adonan tepung ketan yang dicampur dengan pisang raja atau ubi, kemudian diisi dengan kelapa parut manis atau selai serikaya. Adonan tersebut dibungkus menggunakan daun pisang muda yang telah dipanaskan agar lentur dan menghasilkan aroma khas saat dikukus. Proses pembuatan ini mencerminkan kearifan lokal dan keterampilan masyarakat Aceh yang diwariskan secara turun-temurun.

Dalam sebuah usaha rumahan, diproduksi dua jenis timphan, yaitu timphan isian srikaya dan timphan isian kelapa. Untuk memproduksi 50 buah timphan isian srikaya, diperlukan biaya bahan baku sebesar Rp100.000, sehingga harga bahan baku per satuan timphan isian srikaya adalah Rp2.000 per buah. Sementara itu, untuk memproduksi 100 buah timphan isian kelapa, diperlukan biaya bahan baku sebesar Rp100.000, sehingga harga bahan baku per satuan timphan isian kelapa juga Rp1.000 per buah.

1. Berdasarkan data biaya bahan baku timphan isian srikaya dan timphan isian kelapa, perbandingan harga bahan baku timphan isian srikaya : timphan isian kelapa dalam bentuk paling sederhana adalah....

- A. 1:1
- B. 2:1
- C. 1:2
- D. 4:2

Figure 2. Cultural Context Item Number 1

The test reliability index calculated using Cronbach's Alpha coefficient was 0.712, indicating that the instrument has high internal consistency. According to Masrun (1979), a test is considered reliable if its reliability coefficient is at least 0.70. This finding is also consistent with the study by (Dany & Duskri, 2026), which found that instruments with high reliability are capable of producing consistent and accurate measurements of students' abilities. In addition, the Standard Error of Measurement (SEM) value of 1.599 indicates that the level of measurement error is relatively low, so the test results can be trusted to represent students' mathematical literacy abilities.

In terms of difficulty level, the composition of the items consisted of one easy item (7.7%), eleven moderate items (84.6%), and one difficult item (7.7%). This composition indicates that the instrument is dominated by items with a moderate level of difficulty, which is ideal for measuring students' abilities accurately. In terms of discrimination index, ten items (76.9%) were classified as very good, two items (15.4%) were classified as good, and one item (7.7%) was classified as fair, with no items showing poor discrimination. These findings indicate that, overall, the instrument is capable of distinguishing students based on their level of mathematical literacy ability.

Distractor Analysis

The analysis of distractor functioning for the 13 multiple-choice items (each with 3 distractors) showed very good results, as presented in Table 9.

Tabel 9. Distractor Effectiveness Analysis

Item No.	Alternative (Key)	Proportion of Effective Distractors	Number of Effective Distractors	Category
1	B	A (11.7%), C (15.6%), D (2.3%)	2 out of 3	Good
2	C	A (31.3%), B (29.7%), D (5.5%)	3 out of 3	Very Good
3	D	A (24.2%), B (16.4%), C (17.2%)	3 out of 3	Very Good
4	C	A (9.4%), B (15.6%), D (8.6%)	3 out of 3	Very Good

Item No.	Alternative (Key)	Proportion of Effective Distractors	Number of Effective Distractors	Category
5	C	A (21.9%), B (10.2%), D (16.4%)	3 out of 3	Very Good
6	B	A (14.1%), C (20.3%), D (22.7%)	3 out of 3	Very Good
7	D	A (15.6%), B (21.1%), C (30.5%)	3 out of 3	Very Good
8	A	B (14.8%), C (21.9%), D (17.2%)	3 out of 3	Very Good
9	A	B (20.3%), C (20.3%), D (16.4%)	3 out of 3	Very Good
10	B	A (29.7%), C (14.8%), D (16.4%)	3 out of 3	Very Good
11	D	A (14.8%), B (15.6%), C (18.0%)	3 out of 3	Very Good
12	C	A (30.5%), B (21.9%), D (20.3%)	3 out of 3	Very Good
13	A	B (18.8%), C (18.8%), D (14.8%)	3 out of 3	Very Good

Based on Table 9, distractor effectiveness analysis was conducted to determine the ability of each incorrect answer option to attract students who had not yet mastered the material. The results presented in Table 9 show that most distractors functioned very well. Of the total 39 distractors analyzed, 38 distractors (97.44%) were selected by more than 5% of the test participants and therefore met the criteria for effective distractors. Only one distractor did not function optimally, namely option D in item number 1, which was selected by only 2.3% of the students.

In general, 12 out of 13 items (92.30%) had three effective distractors, while 1 item (7.7%) had two effective distractors. This condition indicates that the developed answer alternatives were able to represent the errors or misconceptions that students might make when solving mathematical literacy items on algebra content. The high percentage of distractor selection in several items, such as items 2, 7, and 12, indicates that the distractors were sufficiently logical and capable of attracting students who had not yet understood the concepts correctly.

Item number 1 was the only item that had one ineffective distractor. This finding is consistent with the results of the difficulty index analysis, which showed that item number 1 was classified as easy ($P = 0.703$). The high proportion of students who answered correctly resulted in only a few students selecting one of the distractors, causing option D to function less effectively as a distractor. Nevertheless, the other two distractors still functioned well, so the item was retained without revision.

Meanwhile, item number 12, which had the highest difficulty level, showed that all distractors functioned effectively, with each distractor being selected by more than 20% of the students. This finding indicates that although the item was classified as difficult, the answer alternatives were able to represent various possible student errors effectively. Therefore, the quality of the distractors in item number 12 was considered to support the item's ability to distinguish different levels of students' mathematical literacy abilities.

Overall, the results of the distractor analysis indicate that the developed answer alternatives fulfilled their function as effective distractors. This finding strengthens the results of the validity, reliability, difficulty index, and discrimination index analyses, which indicate that the instrument has good psychometric quality and is suitable for measuring students' mathematical literacy abilities. Descriptive statistics of students' total scores in the field test are presented in Table 10.

Tabel 10. Descriptive Statistics of Total Score (n= 128)

Statistic	Value
Number of Items (N of Items)	13
Number of Examinees (N of Examinees)	128
Mean Score (Mean)	5.945
Variance	8.880
Standard Deviation (Std. Dev.)	2.980
Reliability (Alpha)	0.712
Mean Difficulty Index (Mean P)	0.457
Mean Point-Biserial Correlation (Mean Item-Tot.)	0.472
Mean Biserial Correlation (Mean Biserial)	0.602

The score distribution showed a mean of 5.945 out of a maximum score of 13, with a standard deviation of 2.980, indicating a fairly wide variation in students' abilities. The score range from 0 to 13 confirms that there were no excessively extreme items, while the skewness value close to zero (0.042) indicates a distribution that approximates normality.

The findings of the field test were supported by student questionnaire data, which were used as part of the instrument practicality evaluation. The questionnaire was administered to 23 students selected purposively after the implementation of the field test to obtain information regarding the attractiveness, level of difficulty, and usefulness of the instrument. The number of respondents was considered adequate because the questionnaire functioned as supporting data for evaluating the practicality and acceptability of the instrument, rather than for testing its psychometric quality, which had already been analyzed using the field test data from 128 students (Tessmer, 1993). A summary of the questionnaire results is presented in Table 11.

Table 11. Summary of Student Questionnaire Results by Aspect

No.	Aspect	Percentage	Category
1	Attractiveness of the Items	71.1%	Good
2	Difficulty Level	55.6%	Moderate
3	Usefulness of the Items	79%	Good

Based on Table 11, the attractiveness aspect obtained a percentage of 71.1% (good category), the difficulty level aspect obtained 55.6% (moderate category), and the usefulness aspect obtained 79% (good category). These results indicate that the developed instrument not only meets the criteria of empirical quality but also has good practicality and acceptability from the students' perspective. Therefore, the questionnaire data further support the feasibility of the instrument for use as a mathematical literacy assessment tool on algebra content for eighth-grade junior secondary school students.

Discussion

The first major finding is that the developed assessment achieved strong expert-based evidence of content and contextual appropriateness, with an overall validation score of 4.40 and particularly strong judgments for the relevance of the Acehese cultural context. This result is important because contextualization in mathematical literacy assessment is defensible only when the context supports, rather than obscures, the

mathematical reasoning that students are expected to demonstrate, particularly the processes of formulating, employing, and interpreting mathematics described in the PISA framework (OECD, 2023). The high appraisal of contextual relevance suggests that incorporating timphan, Gayo coffee, meugang, and other locally recognizable situations did not merely decorate conventional algebra problems but provided plausible settings in which quantitative relationships could be interpreted. This interpretation is consistent with culturally relevant mathematics research showing that students' sociocultural experiences can serve as legitimate resources for connecting formal mathematical knowledge with meaningful situations (Kolovou, 2023; Kabuye Batiibwe, 2024; Fitriza et al., 2024). It also extends the literature beyond studies that primarily employ a single ethnomathematical setting by showing that local cultural situations can coexist with individual contexts within one assessment while preserving a common algebraic literacy construct, an issue that remains less developed in previous contextual assessment research (Fouze & Amit, 2025; Legarde, 2026). At the same time, the relatively lower expert ratings for language clarity and visual appropriateness indicate that cultural familiarity alone does not guarantee assessment accessibility; contextual authenticity must be accompanied by precise wording and functional visual representation. Thus, the expert-review evidence supports the theoretical premise of culturally responsive assessment while refining it by indicating that context should be understood as a construct-relevant design resource whose effectiveness depends on its alignment with mathematical content, linguistic clarity, and the intended literacy processes.

The second major finding concerns the empirical functioning of the 13 algebra-based mathematical literacy items, all of which remained acceptable in the field test while displaying meaningful variation in difficulty and discrimination. This pattern indicates that contextualization did not prevent the items from differentiating among students with different levels of mathematical literacy, which is particularly important because algebraic literacy requires movement between contextual information and symbolic relationships rather than procedural calculation alone. The finding is theoretically consistent with the PISA conception of mathematical literacy, in which successful performance requires students to formulate situations mathematically, employ relevant mathematical knowledge, and interpret results in relation to the original context (OECD, 2023), and it accords with research emphasizing the importance of contextual reasoning for meaningful mathematical problem solving (Koskinen & Pitkäniemi, 2022; Çakıroğlu et al., 2024). It also complements recent assessment-oriented work showing that algebraic and literacy constructs require defensible empirical item characteristics rather than relying solely on judgments about content relevance (Chen et al., 2023; D'Erchie et al., 2026). Particularly informative is the contrast between Item 4, which combined moderate difficulty with the strongest discrimination, and Item 12, which was difficult but retained acceptable discriminatory capacity; this suggests that greater contextual or reasoning demand does not necessarily make an item psychometrically weak. Instead, the more difficult item may require a more complex coordination of contextual information and algebraic representation, an interpretation compatible with studies emphasizing reasoning and nonroutine problem solving as central features of meaningful mathematical assessment

(Nilimaa, 2023; Choi et al., 2026). The contribution of the present findings is therefore not simply that contextual algebra items can be developed, but that individual and culturally situated problems can function across different levels of cognitive demand while remaining empirically informative about students' mathematical literacy.

A third important finding is the convergence of the item-level evidence with acceptable test-level internal consistency, reflected in an alpha coefficient of .712, a mean difficulty index of .457, and a mean item-total correlation of .472. Rather than interpreting reliability in isolation, these statistics indicate that the 13 items functioned sufficiently coherently as an assessment while retaining variation in the challenges they presented to students. Such a balance is important because an assessment of mathematical literacy should not consist of highly homogeneous procedural items merely to maximize internal consistency; it must sample different situations and reasoning demands while continuing to measure a defensible common construct. This interpretation aligns with recent instrument-development literature emphasizing the need to establish measurement quality when complex educational constructs are operationalized through test items (Chen et al., 2023; Hwang, 2025; D'Erchie et al., 2026). The predominance of moderately difficult items, together with very good or good discrimination for 12 of the 13 items, further suggests that contextual diversity did not produce substantial construct fragmentation, despite the instrument drawing on both personal situations and Acehnese cultural settings. In this respect, the findings extend previous algebra-oriented assessment work by indicating that psychometric coherence can be maintained when contextual diversity is intentionally designed at the instrument level rather than restricted to a single category of situation. Nevertheless, the reliability coefficient should be interpreted as evidence of acceptable internal consistency within the present sample rather than as proof of universal measurement stability, making replication across populations and more advanced psychometric modeling an important next step.

The fourth finding, that 38 of the 39 distractors (97.44%) functioned effectively, provides additional evidence that the instrument did more than distinguish correct from incorrect responses; most alternatives were sufficiently plausible to represent possible errors in students' algebraic reasoning. This is significant for contextual mathematical literacy assessment because a distractor is informative only when it reflects a response that students might reasonably select after misinterpreting a relationship, applying an inappropriate procedure, or failing to coordinate contextual information with an algebraic model. The pattern is consistent with the broader argument that meaningful mathematics assessment should elicit reasoning and expose differences in how learners interpret and use mathematical information rather than merely reward routine recall (Nilimaa, 2023; Choi et al., 2026). Item 1 provides a useful counterexample: it was the easiest item and contained the only distractor selected by fewer than 5% of students, suggesting that an accessible item can reduce the plausibility of some incorrect alternatives because the correct solution is more readily recognized. By contrast, all distractors in Item 12, the most difficult item, attracted more than 20% of students, indicating that greater task complexity generated several plausible pathways to incorrect responses rather than random failure. This relationship between item difficulty and distractor functioning strengthens the

interpretation that the assessment captured variation in students' reasoning processes, although the present multiple-choice data cannot establish the precise misconceptions underlying each incorrect choice without complementary cognitive interviews or response-process evidence. Consequently, the distractor findings add an important layer to the psychometric evidence by showing that most incorrect options functioned as meaningful components of the assessment rather than as obviously implausible alternatives.

Finally, the student-response evidence indicates that the instrument was not only empirically acceptable but also reasonably usable from the perspective of its intended users, with attractiveness rated at 71.1%, usefulness at 79%, and perceived difficulty at a more moderate 55.6%. The combination is theoretically meaningful because a mathematical literacy assessment should make contextual problems sufficiently interpretable and relevant without removing the intellectual challenge required to formulate, employ, and interpret mathematics in nonroutine situations (OECD, 2023). The favorable ratings for attractiveness and usefulness are consistent with research arguing that culturally and contextually meaningful mathematics can strengthen connections between formal mathematical ideas and learners' lived experiences (Fitriana et al., 2024; Kabuye Batiibwe, 2024; Asare, 2026), while the moderate difficulty perception cautions against equating contextual familiarity with cognitive simplicity. This distinction is central to the novelty of the study: individual experiences and Acehnese cultural situations were not combined within each item, but were deliberately distributed across a single algebra assessment, allowing contextual diversity to operate at the instrument level while preserving a shared mathematical literacy orientation. In this respect, the study bridges two strands that have often developed separately in the literature, namely psychometrically oriented mathematical literacy assessment and culturally grounded mathematics tasks, thereby extending culturally responsive perspectives from the use of local examples toward systematic contextual diversification in assessment design (Kolovou, 2023; Banjo et al., 2025; Fouze & Amit, 2025). Practically, this configuration provides teachers with a model for assessing algebra through situations connected both to students' personal lives and their sociocultural environment, while theoretically it suggests that contextual diversity can coexist with acceptable reliability, discrimination, difficulty, and distractor functioning when the contexts are deliberately aligned with the intended construct. These conclusions should nevertheless remain bounded by the four participating schools and the classical item-analysis evidence used here; broader validation across regions, cultural groups, and psychometric frameworks is required before claiming that the instrument functions equivalently beyond the present Grade VIII population in Banda Aceh.

Implications

The findings have important implications for both mathematical literacy assessment and classroom practice at the junior secondary level. The successful development of 13 algebra-based items incorporating individual and Acehnese cultural contexts indicates that contextual diversity can be incorporated within a single assessment without undermining acceptable item functioning, reliability, or discrimination. This supports the use of

assessment designs that connect algebraic reasoning with students' personal experiences and local cultural environments while remaining aligned with the processes of formulating, employing, and interpreting mathematics. For teachers, the instrument can serve as an alternative means of identifying how students respond to algebraic problems presented through different contextual situations, rather than relying exclusively on decontextualized procedural tasks. More broadly, the study provides a practical model for developing locally responsive mathematical literacy assessments that combine contextual relevance with empirical evaluation of item quality.

Limitations

Several limitations should be considered when interpreting the findings of this study. The field test involved 128 Grade VIII students from four schools in Banda Aceh, meaning that the psychometric evidence reflects the characteristics of this particular sample and should not be generalized automatically to students in other regions or cultural settings. The empirical evaluation relied primarily on classical item analysis, so the stability of item parameters across different ability groups or populations was not examined using more advanced approaches such as item response theory. In addition, the student response questionnaire involved only 23 purposively selected participants and therefore served as supporting evidence of usability and acceptability rather than a comprehensive measure of practicality. Finally, because the study focused on instrument development rather than intervention effects, the findings cannot be interpreted as evidence that the use of individual or Acehnese cultural contexts directly improves students' mathematical literacy performance.

Suggestions

Future research should extend the validation of the instrument to larger and more diverse student populations across different regions to examine whether its item characteristics remain stable beyond the Banda Aceh context. Subsequent studies may also apply item response theory or other advanced psychometric approaches to investigate item difficulty, discrimination, and measurement precision across different levels of student ability. The contextual design can be expanded by incorporating cultural settings from other regions of Indonesia while maintaining a clear distinction between contextual relevance and the mathematical construct being assessed. Further research should also examine students' reasoning processes through cognitive interviews, written-response analysis, or think-aloud protocols to identify why particular distractors attract certain response patterns. In addition, longitudinal or experimental studies may be conducted separately to determine whether repeated exposure to contextualized algebra tasks contributes to the development of mathematical literacy, an effect that was not tested in the present study.

CONCLUSIONS

This study developed a 13-item algebra-based mathematical literacy assessment for Grade VIII students that incorporated individual contexts and Acehnese cultural contexts across different items within a single instrument. The development process combined

expert review, one-to-one evaluation, small-group evaluation, and a field test involving 128 students from four schools in Banda Aceh. Expert validation indicated that the instrument had strong content and contextual appropriateness, with particularly high ratings for the relevance of the Acehese cultural context, while language clarity and visual presentation required only minor refinement. Empirical analysis showed that all 13 items were retained, with point-biserial coefficients ranging from 0.299 to 0.609 and an alpha reliability coefficient of 0.712. The instrument was dominated by moderately difficult items, while most items demonstrated good to very good discrimination and 38 of the 39 distractors functioned effectively. These results indicate that the developed assessment possesses acceptable psychometric characteristics for measuring algebraic mathematical literacy within the participating Grade VIII sample.

The main contribution of the study lies in the deliberate representation of two forms of contextual experience, namely students' individual situations and Acehese cultural settings, within one algebra-based mathematical literacy assessment. Rather than combining both contexts within the same item, the instrument distributed them across five individual-context items and eight Acehese cultural-context items, thereby broadening the contextual range of the assessment while preserving a common mathematical literacy focus. The student-response data further showed that the instrument was perceived as attractive and useful, while its level of difficulty remained moderate, providing complementary evidence of usability alongside the psychometric findings. Taken together, the results support the feasibility of using contextual diversity as a design principle in algebra assessment when it is accompanied by systematic expert judgment and empirical item analysis. The findings therefore position the instrument as a context-sensitive assessment tool that connects formal algebraic reasoning with situations drawn from students' personal and sociocultural environments. Within the scope of the present sample and analysis, the study demonstrates that contextual relevance and empirical measurement quality can be maintained simultaneously in the development of mathematical literacy assessment items.

AUTHOR CONTRIBUTIONS STATEMENT

Alya Fadhila conceived the study, conducted the literature and curriculum analyses, developed the mathematical literacy assessment items, coordinated the data collection, performed the empirical item analysis, and prepared the original manuscript draft. M. Duskri supervised the overall research process, provided methodological and conceptual guidance, contributed to the validation and refinement of the assessment instrument, interpreted the findings, and critically reviewed the manuscript for important intellectual content. Both authors contributed to the development of the research design, evaluation of the instrument, interpretation of the results, and revision of the manuscript. Both authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

REFERENCES

- Ammer, M. A., & Aldhyani, T. H. H. (2022). An investigation into the determinants of investment awareness: Evidence from the young Saudi generation. *Sustainability*, 14(20), 13454. <https://doi.org/10.3390/su142013454>
- Asare, B. (2026). Influence of ethnomathematics-based instruction on students' attitudes, participation, and cultural connections in mathematics learning in Ghana. *Cogent Education*, 13(1), 2612395. <https://doi.org/10.1080/2331186X.2025.2612395>
- Banjo, B. O., Luneta, K., & Rodríguez-Nieto, C. A. (2025). Exploring teachers' perceptions of mathematical connections in the mathematics curriculum and teaching. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(9), em2703. <https://doi.org/10.29333/ejmste/16883>
- Çakıroğlu, Ü., Güler, M., Dündar, M., & Coşkun, F. (2024). Virtual reality in realistic mathematics education to develop mathematical literacy skills. *International Journal of Human-Computer Interaction*, 40(17), 4661–4673. <https://doi.org/10.1080/10447318.2023.2219960>
- Chen, Q., Zheng, H., Fan, H., & Mo, L. (2023). Construction of a reading literacy test item bank for fourth graders based on item response theory. *Frontiers in Psychology*, 14, 1103853. <https://doi.org/10.3389/fpsyg.2023.1103853>
- Choi, K. M., Mun, J., & Hand, B. (2026). Bridging reasoning in mathematics and science: A document analysis of applying and reasoning in the TIMSS assessment framework and the case for utilization of resources. *Educational Psychology Review*, 38(1), 22. <https://doi.org/10.1007/s10648-025-10115-2>
- D'Erchie, M., Forsmann, C. L., Schneider, M., & Obersteiner, A. (2026). Discriminant validity and interrelations of conceptual and procedural knowledge in fractions and algebra: Evidence from confirmatory factor analysis. *British Journal of Educational Psychology*, 96(3), 1396–1418. <https://doi.org/10.1111/bjep.70091>
- Dickson, T. (2025). *Integrating goal-setting into an urban community college developmental mathematics course: A multi-perspective learning experience* [Doctoral dissertation]. <https://doi.org/10.70013/g3j2x8vp>
- Elmalı, F., & Çıtak, H. (2026). Design, development, and formative evaluation of a tablet-based augmented reality application for early number learning. *International Journal of Human-Computer Interaction*, 1–21. <https://doi.org/10.1080/10447318.2026.2622591>
- Fitriza, R., Hurriyah, H., & Hadaina, D. (2024). Ethno-mathematics in Minangkabau indigenous society. *Mosharafa: Jurnal Pendidikan Matematika*, 13(3), 661–674. <https://doi.org/10.31980/mosharafa.v13i3.2154>
- Fouze, A. Q., & Amit, M. (2025). The impact of cultural context and ethnomathematics on mathematics education. *Creative Education*, 16(12), 2140–2155. <https://doi.org/10.4236/ce.2025.1612129>
- Hwang, J. (2025). Developing an instrument to measure Korean pre-service teachers' understanding of language as an epistemic tool in mathematics education. *School Science and Mathematics*, 125(6), 600–612. <https://doi.org/10.1111/ssm.18309>
- Kabuye Batiibwe, M. S. (2024). The role of ethnomathematics in mathematics education: A literature review. *Asian Journal for Mathematics Education*, 3(4), 383–405. <https://doi.org/10.1177/27527263241300400>
- Kaya, O. N., & Kaya, Z. (2024). A co-design based research study: Developing formative assessment practices with preservice science teachers in a chemistry laboratory setting. *Research in Science Education*, 54(4), 739–774. <https://doi.org/10.1007/s11165-024-10162-9>

- Kolovou, M. (2023). Embracing culturally relevant education in mathematics and science: A literature review. *The Urban Review*, 55(1), 133–172. <https://doi.org/10.1007/s11256-022-00643-4>
- Koskinen, R., & Pitkäniemi, H. (2022). Meaningful learning in mathematics: A research synthesis of teaching approaches. *International Electronic Journal of Mathematics Education*, 17(2), em0679. <https://doi.org/10.29333/iejme/11715>
- Legarde, M. A. (2026). Reimagining mathematics education through context and culture: Emerging opportunities and pedagogical challenges. *Journal of Qualitative Research in Education*, 46, 129–144. <https://doi.org/10.54963/jqre.i46.2307>
- Monteiro, V., & Passarinho, B. B. (2026). Formative assessment and self-regulated learning in lower secondary mathematics: Students' and teachers' perceptions. *Education Sciences*, 16(3), 452. <https://doi.org/10.3390/educsci16030452>
- Nilimaa, J. (2023). New examination approach for real-world creativity and problem-solving skills in mathematics. *Trends in Higher Education*, 2(3), 477–495. <https://doi.org/10.3390/higheredu2030028>
- Nolan, K., & Xenofontos, C. (2023). Mapping teachers' perspectives on culturally responsive pedagogy in mathematics: From academic achievement to insights and opportunities. *Intercultural Education*, 34(6), 550–567. <https://doi.org/10.1080/14675986.2023.2265855>
- Nurgabyl, D., Satkulov, B., & Kagazbayeva, A. (2023). Formation and development of mathematical literacy in the context of evaluative-study tasks of PISA. *Journal on Mathematics Education*, 14(4), 701–722.
- OECD. (2023). *PISA 2022 results: Country notes: Indonesia*. Organisation for Economic Co-operation and Development.
- Pratiwi, S. A., Peni, N. R. N., & Prabowo, A. (2024). Study on literacy numeracy towards students' logic mathematics: A literature review. *Numeracy*, 11(1), 58–69. <https://doi.org/10.46244/numeracy.v11i1.2601>
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., AlGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(1), 445. <https://doi.org/10.1186/s12909-024-05433-y>
- Schexnayder, J., Perry, K. R., Sheahan, K., Majette Elliott, N., Subramaniam, S., Strawbridge, E., Webel, A. R., Bosworth, H. B., & Gierisch, J. M. (2023). Team-based qualitative rapid analysis: Approach and considerations for conducting developmental formative evaluation for intervention design. *Qualitative Health Research*, 33(8–9), 778–789. <https://doi.org/10.1177/10497323231167348>
- Shah, S. S., & Shah, T. (2023). Responsible consumption choices and individual values: An algebraic interactive approach. *Mind & Society*, 22(1), 1–32. <https://doi.org/10.1007/s11299-023-00294-2>
- Sorby, S. A., Duffy, G., & Yoon, S. Y. (2022). Math instrument development for examining the relationship between spatial and mathematical problem-solving skills. *Education Sciences*, 12(11), 828. <https://doi.org/10.3390/educsci12110828>
- Utami, H. B. (2025). Developing algebra-based test items for measuring students' mathematical literacy. *EduMa: Mathematics Education Learning and Teaching*, 14(1), 330–343. <https://doi.org/10.24235/eduma.v14i1.21581>