Beta: Jurnal Tadris Matematika

Tadris Matematika, Universitas Islam Negeri Mataram, Jl. Gajah Mada No. 100 Jempong, Mataram, Indonesia.

LETTER OF ACCEPTANCE FOR SCIENTIFIC ARTICLE PUBLICATION

Number: 782/Beta/INV/2026

Authors:

Novia Aulia, M. Duskri

Universitas Islam Negeri Ar-Raniry, Aceh, Indonesia, Email:
220205022@student.ar-raniry.ac.id

Title of Article:

Developing mathematical literacy test items in geometry domain using individual and occupational contexts

Congratulations! The editorial Executive Board has evaluated the article through the peer-review process, and it has been "Accepted" and will be published in Beta: Jurnal Tadris Matematika (P-ISSN 2085-5893, E-ISSN 2541- 0458) Vol. 19 No. 2, November 2026.

Thank you for submitting your article to our journal. We wish you all possible success in the future.

Best Regards,

Afifurrahman
Chief Editor

Beta: Jurnal Tadris Matematika, 11(2) 2018: 1-12
DOI 10.20414/betajtm.v9i2.xxx

Research articles

Development of mathematical literacy test items in geometry content domain using individual and occupational contexts for junior high school (SMP/MTs) students

Novia Aulia[1], M. Duskri[2]

Abstrak Penelitian ini bertujuan untuk menghasilkan soal literasi matematis konten geometri dengan konteks individual dan *occupational* yang memenuhi kriteria validitas dan reliabilitas untuk siswa kelas IX SMP/MTs berdasarkan model evaluasi formatif Tessmer, serta mendeskripsikan karakteristik butir soal yang meliputi tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Prosedur pengembangan melalui tahap *preliminary*, *self-evaluation*, *prototyping* (*expert review*, *one-to-one*, *small group*), dan *field test*. Instrumen yang dikembangkan berupa 15 butir soal pilihan ganda yang mencakup level kognitif C2 hingga C5. Pada tahap *expert review*, seluruh butir soal dinyatakan sangat valid oleh tiga validator dengan rata-rata persentase validasi sebesar 87% (kategori sangat valid). Pada tahap *one-to-one* dan *small group*, dilakukan perbaikan redaksional dan teknis berdasarkan masukan siswa, dengan koefisien reliabilitas sebesar 0,744 (kategori tinggi) dan seluruh butir soal dinyatakan valid. Pada tahap *field test* terhadap 130 siswa kelas IX dari empat sekolah di Banda Aceh, diperoleh reliabilitas tes sebesar 0,770 (kategori tinggi). Validitas empiris menunjukkan nilai *point-biserial* antara 0,340 hingga 0,714, yang berarti seluruh butir valid. Distribusi tingkat kesukaran terdiri atas 2 butir mudah, dan 13 butir sedang. Daya pembeda seluruh butir tergolong baik dengan nilai koefisien biserial di atas 0,40. Analisis efektivitas pengecoh menunjukkan bahwa 9 butir soal (60%) berkategori sangat baik dan 6 butir soal (40%) berkategori baik, meskipun empat pengecoh pada butir tertentu perlu direvisi karena dipilih oleh kurang dari 5% peserta uji coba. Dengan demikian, soal literasi matematis yang dikembangkan layak digunakan untuk mengukur kemampuan literasi matematis siswa kelas IX SMP/MTs.

Kata kunci *Literasi Matematis, Geometri, Konteks Individual, Konteks Occupational, Pengembangan*

Abstract This study aims to develop geometry-based mathematical literacy test items with individual and occupational contexts that meet validity and reliability criteria for ninth-grade junior high school/MTs students based on Tessmer's formative evaluation model, as well as to describe the characteristics of the items, including difficulty level, discriminating power, and distractor effectiveness. The development procedure involved the following stages: preliminary, self-evaluation, prototyping (expert review, one-to-one, small group), and field test. The instrument developed consists of 15 multiple-choice items covering cognitive levels from C2 to C5. In the expert review stage, all items were declared very valid by three validators with an average validation percentage of 87% (very valid category). In the one-to-one and small group stages, editorial and technical revisions were made based on student feedback, with a reliability coefficient of 0.744 (high category) and all items declared valid. In the field test stage involving 130 ninth-grade students from four schools in Banda Aceh, the test reliability was 0.770 (high category). Empirical validity showed point-biserial values ranging from 0.340 to 0.714, indicating that all items were valid. The difficulty level distribution consisted of 2 easy items, and 13 moderate items. The discriminating power of all items was classified as good, with biserial coefficient values above 0.40. Distractor effectiveness analysis showed that 9 items (60%) were in the very good category and 6 items (40%) were in the good category, although four distractors on specific items require

[1] Universitas Islam Negeri Ar-Raniry, Aceh, Indonesia, 220205022@student.ar-raniry.ac.id
[2] Universitas Islam Negeri Ar-Raniry, Aceh, Indonesia, m.duskri@ar-raniry.ac.id

revision as they were selected by fewer than 5% of test participants. Thus, the mathematical literacy items developed are suitable for measuring the mathematical literacy skills of ninth-grade junior high school and MTs students.

Keywords *Mathematical Literacy, Geometry, Individual Context, Occupational Context, Development Research*

Introduction

Education is a fundamental pillar for the advancement of a nation, as a high-quality educational process enables individuals to develop their full potential optimally. Law Number 20 of 2003 concerning the National Education System emphasizes that education is a conscious effort to create learning processes that allow students to develop their abilities and potential to the fullest extent (Syafni, 2017). One of the disciplines that plays a significant role in education is mathematics, which contributes to the development of rational and systematic thinking. The National Council of Teachers of Mathematics (NCTM, 2000) emphasizes that effective mathematics instruction should foster students' ability to solve contextual problems. This view is consistent with the definition of mathematical literacy proposed by the Organisation for Economic Co-operation and Development (OECD), which defines mathematical literacy as an individual's capacity to formulate, employ, and interpret mathematics in a variety of real-life contexts (OECD, 2019). In this study, the characteristics of mathematical literacy items are based on the indicators proposed by Ayuningtyas (2017), namely: (1) identifying information and mathematical forms from real-world problems, (2) transforming information from a problem situation into an appropriate mathematical model, (3) using the mathematical model to solve the problem, and (4) interpreting and evaluating the reasonableness of the obtained results within real-life contexts.

In the era of globalization and the Fourth Industrial Revolution, mathematical literacy has become one of the essential twenty-first-century skills, encompassing not only conceptual understanding but also critical thinking, creativity, collaboration, and communication (Trilling & Fadel, 2009). This competency extends beyond mastery of mathematical concepts to include the ability to think critically, generate innovative ideas, collaborate effectively, and communicate clearly. In Indonesia, the national curriculum, through the Minimum Competency Assessment (AKM), places numeracy as one of the primary domains assessed, in alignment with the frameworks of PISA and TIMSS (Kemendikbud, 2020). Nevertheless, Indonesian students' achievement in mathematics remains a concern. The results of TIMSS 2015 ranked Indonesia 44th out of 49 participating countries, with an average score of 397, which was substantially below the international average of 500 (Mullis *et al.*, 2016). A similar trend was observed in PISA 2022, where Indonesia ranked 72nd out of 81 countries, achieving an average score of 379 compared with the international average of 489 (OECD, 2023). These findings indicate that Indonesian students generally continue to experience difficulties in solving mathematical literacy tasks, particularly those at Levels 1 and 2 (Masfufah & Afriansyah, 2021). Likewise, the results of the National Examination (UN) for junior secondary school mathematics showed relatively low achievement, with a national average score of 46.56 in 2019 (Kemendikbud, 2019). These findings suggest that Indonesian students still encounter various challenges in applying mathematical concepts across different contexts, especially in situations requiring complex problem-solving abilities. A similar pattern can also be observed in the fluctuating literacy achievement levels reported in several schools in Banda

Aceh, including MTsN Tungkop, MTsN 1 Banda Aceh, and SMP Negeri 8 Banda Aceh, highlighting the need for continuous evaluation and innovation in both instructional strategies and assessment instruments (Kemendikbudristek, 2024).

The low level of mathematical literacy among Indonesian students is not only a national issue but also a matter of global concern. This condition is influenced by various external and internal factors. From an external perspective, the mathematics curriculum in Indonesia tends to emphasize the mastery of concepts and procedures without providing sufficient opportunities for students to apply them in real-life situations. As a result, mathematics instruction remains predominantly teacher-centered, with learning interactions often occurring in a one-way manner (Fauzi & Arifin, 2020). Consequently, students experience difficulties in connecting mathematical ideas to real-world situations because the tasks presented to them are generally procedural and abstract in nature. From an internal perspective, all aspects of mathematical literacy, including communication, modeling, representation, and reasoning, are more readily mastered by students with a high degree of learning independence, whereas those with lower levels of learning independence tend to encounter greater difficulties (Aldiansyah & Agoestanto, 2024). Furthermore, the level of geometric thinking has a significant influence on mathematical literacy skills. Students who have reached the *informal deduction* level (Van Hiele Level 2) have been shown to formulate problems, apply mathematical concepts, and evaluate solutions more effectively than students at the visualization level (Rahmayani *et al.*, 2024). These findings underscore the importance of fostering geometric thinking skills as a primary focus of mathematics instruction, particularly in learning topics that involve abstract concepts.

The learning context plays a crucial role in the development of students' mathematical literacy. Research has shown that students' mathematical abilities vary depending on the contextual domain employed, and they continue to experience difficulties, particularly in interpreting and evaluating mathematical results (Ayutias *et al.*, 2024). This condition highlights the need for instructional approaches that systematically connect mathematical concepts to practical situations. Context-based tasks have emerged as a strategic solution because they are designed to bridge abstract mathematical concepts with concrete real-life situations. Previous studies have demonstrated that relating mathematics problems to contexts that are familiar to students can enhance their motivation and engagement in learning (Sari & Putra, 2021).

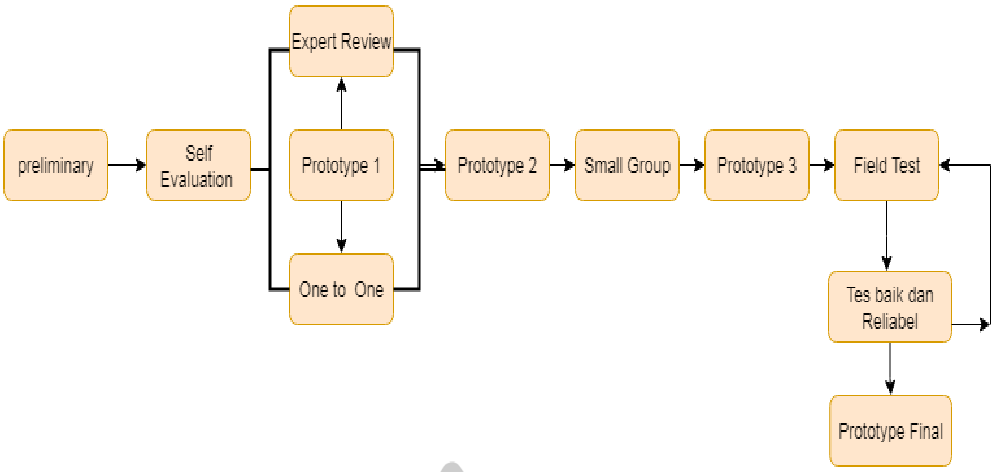
In the teaching of geometry at the ninth-grade level, a context-based approach becomes increasingly relevant, as geometry offers substantial opportunities to be connected with various real-world situations. Individual contexts may be represented through personal problems, such as calculating the area of a bedroom or designing a garden, whereas occupational contexts may be developed through workplace-related situations, such as determining the volume of construction materials or analyzing production costs that are appropriate to the developmental level of junior secondary school students. Nevertheless, geometry instruction continues to face several challenges, including the predominance of theoretical and abstract problems, limited use of visual learning media, and time constraints resulting from a dense curriculum (Fauzi & Arifin, 2020). In addition, students often encounter difficulties in understanding geometric concepts because of their abstract nature and their limited connection to everyday experiences (Wijaya, 2018).

The pedagogical implications of these findings underscore the importance of designing geometry assessment instruments that connect abstract mathematical concepts with real-life situations while meeting psychometric standards in terms of validity, reliability, and practicality. Such instruments can make geometry learning more meaningful and contribute comprehensively to the development of students' mathematical literacy. Based on the literature review, the availability of geometry-based mathematical literacy items that proportionally integrate both individual and occupational contexts for Grade IX junior secondary school students remains very limited. Although the development of context-based tasks has been shown to effectively enhance students' mathematical literacy and learning interest (Isnaniah & Imamuddin, 2022). their average literacy achievement remains relatively low due to difficulties in translating word problems into formal mathematical representations (Hapsari, 2017). Furthermore, geometry problems presented in existing textbooks are still dominated by procedural tasks that have not optimally stimulated higher-order thinking skills (Adinawan, 2016).

Based on this gap, the present study focuses on the development of a mathematical literacy test in the geometry domain incorporating individual and *occupational* contexts for Grade IX junior secondary school students in the form of multiple-choice items. The novelty of this study lies in the design of distractors that are constructed based on common student error patterns, including conceptual errors, principle-related errors, careless mistakes, and misinterpretations of problem statements. Consequently, each selected distractor can function as a diagnostic tool for identifying students' specific learning difficulties (Duskri *et al.*, 2014). This approach distinguishes the present study from previous research, which has generally focused only on validity and reliability without considering the diagnostic function of distractors. Moreover, it addresses a research gap concerning the limited integration of individual and occupational contexts within a single set of geometry literacy items for Grade IX junior secondary school students. The developed instrument is expected to assist teachers in identifying students' abilities and learning difficulties more accurately, thereby enabling students to receive more targeted learning strategies that are aligned with their specific needs.

Methods

This study employed a *Research and Development* (R&D) approach aimed at producing a set of mathematical literacy test items in the geometry content domain using individual and occupational contexts for Grade IX junior secondary school students. The development model adopted in this study was Tessmer's formative evaluation model (Tessmer, 1993), which consists of four main stages: the *preliminary* stage, *self-evaluation*, *prototyping* (including *expert review*, *one-to-one*, and *small group* evaluations), and the *field test* stage. The development of the test items began with the preparation of a test blueprint based on the geometry competencies for Grade IX junior secondary school students, the mathematical literacy frameworks proposed by the OECD and Ayuningtyas, and the revised Bloom's Taxonomy. Each item was developed through several stages, namely determining the indicators, selecting the context (individual or *occupational*), constructing the stimulus, formulating the item stem, developing the answer options, and designing distractors based on common student error patterns. Subsequently, the items were validated by experts and revised according to the feedback obtained during the *expert review*, *one-to-one*, and *small group* stages before being administered in the *field test* stage. The development procedure is presented in Figure 1.



.

Figure 1. Flowcart of Tessmer's Formative Evaluation Model

This study was conducted in several schools in Banda Aceh during the second semester of the 2025/2026 academic year. The participants in the *field test* stage consisted of 130 Grade IX students from four schools, namely SMP Negeri 1 Banda Aceh, SMP Negeri 3 Banda Aceh, SMP Negeri 17 Banda Aceh, and MTsN 1 Banda Aceh. The participants were selected using a *purposive sampling* technique based on the criterion of having moderate academic ability and not belonging to either the high-achieving or low-achieving groups. This criterion was considered important to ensure that the item characteristics obtained, including item difficulty, discrimination index, and item quality indices, could adequately represent the general abilities of junior secondary school students. A sample size of 130 students was deemed sufficient because, in instrument development research, a minimum of 100 respondents is generally considered adequate for producing stable item parameter estimates (Santoso, 2017), although a larger-scale trial could not be conducted due to limitations in time, resources, and funding. After completing the test, 36 students were selected through *purposive sampling* to complete a response questionnaire regarding language clarity, readability, and the appropriateness of the contexts used in the developed items. This number was considered representative because it met the minimum sample adequacy principle for small-scale survey research, which recommends at least 30 respondents (Fraenkel *et al.*, 2012; Sugiyono, 2019).

The development procedure began with the *preliminary* stage, which involved a needs analysis, followed by the *self-evaluation* stage, including curriculum analysis, student analysis, content analysis, and the initial design of the test items (Prototype I). Subsequently, during the *prototyping* stage, the process started with an *expert review* involving three validators: one educational evaluation lecturer who assessed construct validity and conformity with assessment principles, and two mathematics teachers who evaluated the suitability of the geometry content, student characteristics, and mathematical literacy indicators.

The validation instrument used in this study was a validation sheet consisting of four main aspects: content, construction, context, and language (Mulyatiningsih, 2011; Kemendikbud, 2017). The content aspect focused on the alignment of geometry content with the basic competencies, the suitability of the items with mathematical literacy indicators, and the relevance of the individual and occupational contexts. The construction aspect included the clarity of item statements, the accuracy of answer keys, and the functionality of distractors designed based on common student error patterns (Retnawati, 2016). he language aspect emphasized the use of proper and correct language as well as the absence of ambiguity. Each aspect was assessed using a five-point rating scale and was accompanied by a section for suggestions and revisions. The validators' evaluations were then used as the basis for revising *Prototype I* into *Prototype II*.

The developed mathematical literacy test consisted of 15 multiple-choice items, each with four answer options, distributed across cognitive levels C2 to C5. The distribution comprised 2 items at C2 (13.3%), 6 items at C3 (40.0%), 5 items at C4 (33.3%), and 2 items at C5 (13.3%). Cognitive level C1 was not included because the ability to recall basic facts and procedures was assumed to have already been mastered by Grade IX students as a prerequisite for further learning. This distribution was designed to align with the PISA framework, which emphasizes the processes of *formulating*, *employing*, and *interpreting* mathematics (corresponding primarily to C2–C4), as well as the ability to evaluate the reasonableness of solutions in real-life contexts (Wilson, 2016; OECD, 2023)(OECD, 2023). Therefore, the measurement of students' mathematical literacy skills could be conducted progressively in accordance with the characteristics and developmental level of Grade IX junior secondary school students..

Data were collected using two techniques: expert validation to assess the feasibility of the test prototypes and testing at each trial stage to obtain empirical data regarding item quality. The data were then analyzed using descriptive quantitative methods, including validity testing through the *point-biserial correlation coefficient*, reliability testing using *Cronbach's alpha coefficient*, and analyses of item difficulty, discrimination index, and distractor effectiveness.

The instrument validation data obtained from the validators were collected using a Likert scale, as presented in Table 1 (Widyawati & Duskri, 2025).

Table 1. Validaity criteria

Scale	Category	Description
1	Strongy dissagree	Does not meet the criteria
2	Dissagree	Does not meet the criteria
3	Fair	Particially meets the criteria
4	Agree	Meets the criteria
5	Strongly agree	Fully meets the criteria

Each item in the instrument was evaluated by the validators based on the criteria presented in Table 1. The evaluation covered four aspects: content/material, construction, language, and context. The scores obtained were subsequently used to calculate the instrument validation score. The formula used to analyze the validation data is presented as follows.

$$nilai\ validator\ akhir = \frac{score\ obtained}{maximum\ score}\ x\ 100$$

Subsequently, the validation scores were interpreted according to the validity criteria presented in Table 2 (Utami et al., 2025).

Table 2. Validity score criteria

Score interval	Category
$0 - 20$	Very invalid
$21 - 40$	Invalid
$41 - 60$	Less invalid
$61 - 80$	Valid
$81 - 100$	Very valid

A test is considered reliable if its reliability coefficient is at least 0,70 (Masrun, 1979). In educational measurement, such as teacher-made tests, reliability coefficients generally range from 0,60 to 0,85 (Grounlund, 1982). In this study, reliability was interpreted as high when the coefficient ranged from 0,60 to 1,00, as shown in Table 3 (Basri & Turmuzi, 2021).

Table 3. Reliability coeffecient category

Interval score	Category
$0{,}00 \leq r_{11} < 0{,}20$	Very low reliability
$0{,}20 \leq r_{11} < 0{,}40$	Low reliability
$0{,}40 \leq r_{11} < 0{,}60$	Moderate reliability
$0{,}60 \leq r_{11} < 0{,}80$	High reliability
$0{,}80 \leq r_{11} < 1{,}00$	Very high reliability

Item difficulty refers to the proportion of test-takers who answer an item correctly (Suryabrata, 1987). Based on the difficulty index (p), items are considered acceptable when $0{,}30 \leq p \leq 0{,}70$, rejected when $p < 0{,}10$ or $p > 0{,}90$, and revised when $0{,}10 \leq p \leq 0{,}29$ or $0{,}70 \leq p \leq 0{,}90$ (Pakpahan, 1990). In this study, items were considered acceptable if their difficulty index ranged from 0,30 to 0,70 (Bano et al., 2022) as presented in table 4.

Table 4. Item difficulty clasification

Interval Score	Interpretasi
$0{,}1 - 0{,}29$	Difficult, poor, need revision
$0{,}30 - 0{,}70$	Moderate, good item, accepted
$0{,}71 - 0{,}90$	Easy, good item, nees revision

Item discrimination refers to the ability of an item to distinguish between high-performing and low-performing students. According to Crocker & Algina (1986), discrimination coefficients ranging from 0,40 to 1,00 are considered good, 0,30 – 0,39 are acceptable without revision, 0,20 – 0,29 require revision, and values from –1,00 to 0,19 indicate poor or non-functioning items. In this study, items were considered acceptable if they had discrimination indices ranging from 0,30 to 1,00 (Qadir et al., 2024). As shown in Table 5. (Widyawati & Duskri, 2025).

Table 5. Item discrimination clasification

Interval	Category
0,40 – 1,00	Good
0,30 – 0,39	Accepted without revision
0,20 – 0,29	Revision
-1,00 – 0,19	Rejected

Test items consist of two components, namely the item stem and the answer options. The answer options are further divided into the correct answer and distractors. A distractor is considered effective when it is selected more frequently by low-performing students, while high-performing students tend not to choose it. Based on distractor effectiveness criteria, a distractor is regarded as functioning well if it is selected by at least 5% of all test participants. If fewer than 5% of participants choose a distractor, it should be revised, whereas a distractor that is not selected by any participant (0%) should be rejected (*Analisis Butir Soal Dengan BIGSTEPS*, 1997; Arhin, 2024). In line with these criteria, the distractors in this study were designed based on common student error patterns, including conceptual misunderstandings, principle-related errors, computational carelessness, and misinterpretation of contextual information. Distractors selected by fewer than 5% of participants were considered ineffective and recommended for revision.

Findings and Disscusion

The *preliminary* stage aimed to conduct a needs analysis and curriculum analysis to ensure that the developed product was aligned with both curriculum requirements and students' needs. The curriculum analysis focused on the geometry competencies for Grade IX junior secondary school students, including plane figures, solid figures, and angle relationships. Based on this analysis, item indicators were formulated to assess students' ability to apply geometric concepts in solving real-life problems. Geometry instruction at the junior secondary level should indeed be connected to real-world contexts in order to help students develop meaningful understanding (Fauzi & Arifin, 2020).

Student analysis was conducted to identify the characteristics of Grade IX students as the primary users of the developed instrument. Students aged 14–15 years are generally at the formal operational stage according to Piaget's theory of cognitive development, a stage in which they begin to think logically and abstractly, although they still require relevant contexts to support their understanding (Rahayu, 2020). Therefore, the developed items were designed using clear language, contexts closely related to students' daily lives, supporting illustrations, and varying levels of difficulty. In addition to student analysis, a context analysis was conducted to determine the types of contexts to be integrated into the test items. Based on the PISA framework, two types of contexts were selected: individual contexts, which relate to students' personal experiences such as calculating the area of a bedroom or planning wall painting, and occupational contexts, which relate to workplace situations such as calculating the volume of construction materials or determining the area of walls to be covered with tiles. The integration of these two contexts was intended to train students to apply geometric concepts in real-life situations while simultaneously introducing the relevance of mathematics to the world of work (OECD, 2019).

Based on the results of the preliminary stage, Prototype I was developed, consisting of 15 multiple-choice items (7 items on plane figures, 7 items on solid figures, and 1 item on angle relationships) with four answer options. Each item was developed by considering the alignment with competency indicators, the proportional use of individual and occupational contexts, variations in difficulty level, completeness of illustrations, and clarity of language and instructions. Prior to expert validation, the researcher conducted a self-evaluation stage involving a self-review of the alignment between the items and indicators, as well as the clarity of language. According to Tessmer (1993), this stage is essential to ensure that the instrument has a strong theoretical foundation before being tested. Subsequently, Prototype I was validated during the expert review stage.

The *expert review* stage was intended to validate the mathematical literacy items developed by the researcher. The validation process aimed to evaluate the quality of the instrument in terms of content appropriateness, measured competencies, indicators, cognitive levels, language, and item clarity. The validation process involved three validators. The first validator was a lecturer from the Mathematics Education Study Program at UIN Ar-Raniry who had extensive experience in designing mathematics examination items. The second and third validators were mathematics teachers at junior secondary schools who were knowledgeable about both item development and the geometry content used in this study. The assessment employed a five-point Likert scale, where a score of 1 indicated *strongly inappropriate* and a score of 5 indicated *highly appropriate*. The comments and suggestions provided by the validators are presented in Table 6.

Table 6. Validators' comments and sugestions

Item no.	Validator	Comments and suggestions	Revision decision
4	KM	In Passage II, clarify that the mat should completely cover the floor neatly without any joints. Specify the area of the mat that must be ordered.	Added a statement in Passage II indicating that the mat must be ordered as a single piece without joints.
7		Replace the table illustration with one that is genuinely triangular.	Revised the table illustration in Figure 5 using a proportionally accurate equilateral triangle.
10		It would be better to include a figure	Added a visualization showing vertically opposite angles labeled $(4x-20)°$ and $100°$.
11		Add an introductory sentence, for example, "A house...".	Added the introductory statement: "A family is renovating their house..." to the passage
12		Add an introductory story.	Added an introductory paragraph about Roni, an aquarium, and decorative stones in Passage IV.
13		Explain why the package needs to be enlarged.	Added a statement explaining that the package size was increased due to family demand.
14		In Passage V, explain what logistics are, or at least specify the items to be stored	Added examples of logistics goods (rice, sugar, cooking oil, etc.) at the beginning of Passage V
Several item	CA	The language used should not be too advanced.	Simplified the language throughout the items by replacing overly technical terms.
Several item	CA	The texts should not be excessively long.	Shortened the passages by removing non-essential descriptive sentences.
Several item	RF	Some items should include visualizations and explanatory labels.	Added illustrations and captions to items requiring visual support.
Several item	RF	The indicators should more explicitly emphasize mathematical literacy processes.	Revised the indicators to better reflect mathematical literacy processes.

The validation results of the mathematical literacy test on geometry content with individual and occupational contexts for junior secondary school students, based on expert evaluations using a five-point Likert scale assessing content, construction, context, and language aspects, are presented in Table 7.

Table 7. Validation results

Aspec	Presentage	Category
Content	93	Very Valid
Construction	86	Very Valid
Context	90	Very Valid
Language	80	Very Valid
Average	87	Very Valid

At this stage, all 15 test items were deemed suitable for use as a mathematical literacy assessment instrument. The validation results showed scores of 93% for the content aspect, 86% for construction, 90% for context, and 80% for language, resulting in an overall average validity score of 87%, which falls within the *very valid* category.

The *one-to-one* evaluation stage was conducted at MTsN 1 Banda Aceh and involved six Grade IX students with heterogeneous ability levels (high, medium, and low). The participants were selected based on recommendations from mathematics teachers and their previous academic records. This stage aimed to evaluate item readability, clarity of instructions, and students' understanding of the language and illustrations used in the test items, particularly those related to solid geometry, which require visualization skills (Tessmer, 1993).

During the implementation of the *one-to-one* evaluation, students were asked to complete the test individually and were subsequently interviewed to obtain their responses to each item, particularly those they considered difficult. The qualitative analysis indicated that, in general, the students were able to understand the intended meaning and language used in the test items. However, several aspects required revision. Specifically, Items 5 and 12 contained sentence structures that were insufficiently clear and caused confusion among some students. In addition, the illustration used in Item 8 was considered to require improvement in terms of proportionality to enhance comprehensibility. Based on students' feedback at this stage, revisions were made to these items, resulting in *Prototype II*. A summary of the revisions is presented in Table 8.

Table 8. revisions of test item during the one-to-one evaluation stage

Item no.	Before revisions	Findings	Revisions result
5	The item required students to determine the most economical painting cost between two contractors. However, the phrase "*paling ekonomis*" was unclear. In addition, one of the answer options compared a unit cost with a fixed-price contract, making the comparison illogical.	The phrase "*paling ekonomis*" led to different interpretations among students.	The wording was revised to "the lowest total cost," and the answer options were adjusted to ensure logical consistency.
8	The item described storing cube-shaped objects inside a rectangular prism-shaped box. However, the illustration was disproportionate and lacked dimensional information.	The rectangular prism illustration was not proportional and did not	The illustration was redesigned with appropriate proportions and labeled with length, width, and height dimensions.

		provide dimensi	
12	The item described an aquarium that was "*Terisi setengah*," but it was unclear whether this referred to half of the aquarium's volume or half of its height. Furthermore, the characteristics of the object used (stones) were not specified.	The phrase "*Terisi Setengah*" was considered ambiguous.	illustration was redesigned with appropriate proportions and labeled with length, width, and height dimensions.

After revisions were made based on students' feedback, all test items were considered suitable to proceed to the *small-group* stage (Fraenkel *et al.*, 2012).

The *small-group* evaluation was conducted at SMP Negeri 1 Banda Aceh and involved 31 Grade IX students. The school was selected based on its accreditation status, accessibility, and its representation of students with diverse ability levels. The purpose of this stage was to obtain preliminary data regarding item quality before conducting a larger-scale field test and to identify items that still required revision or replacement, particularly those related to solid geometry and angle relationships, which had previously shown relatively high levels of difficulty. The results of the item analysis conducted during the *small-group* stage are presented in Table 9.

Table 9. Item analysis results in the small groups stage

Item no.	Content area	Difficulty index	Category	Discrimination index	category	Validity
1	Plane figures	0,67	Moderate	0,32	Fair	Valid
2	Plane figures	0,61	Moderate	0,45	Good	Valid
3	Solid figures	0,67	Moderate	0,32	Fair	Valid
4	Plane figures	0,61	Moderate	0,45	Good	Valid
5	Plane figures	0,35	Moderate	0,32	Fair	Valid
6	Solid figures	0,58	Moderate	0,38	Fair	Valid
7	Plane figures	0,35	Moderate	0,32	Fair	Valid
8	Solid figures	0,58	Moderate	0,38	Fair	Valid
9	Solid figures	0,35	Moderate	0,45	Good	Valid
10	Angle relationship	0,45	Moderate	0,38	Fair	Valid
11	Solid figures	0,32	Moderate	0,38	Fair	Valid
12	Solid figures	0,51	Moderate	0,51	Good	Valid
13	Solid figures	0,25	Difficult	0,25	Fair	Valid
14	Plane figures	0,38	Moderate	0,38	Fair	Valid
15	Solid figures	0,29	Difficult	0,58	Good	Valid

Note: r-table for n = 31 is 0,355

Based on the results of the *small-group* analysis, all 15 test items were declared valid because the obtained *point-biserial* correlation coefficients ranged from 0,359 to 0,650 and exceeded the critical *r*-table value of 0,355 (α = 0,05; df = 29). According to Retnawati Retnawati (2016), n item is considered valid when its *point-biserial* correlation coefficient is greater than the corresponding *r*-table value. These findings indicate that each item contributed significantly to the mathematical literacy construct being measured. This result is consistent with the findings of Fadilah & Sutrisno (2021) who reported that pilot testing involving 30–50 students is generally sufficient to produce stable validity estimates..

Test reliability was calculated using *Cronbach's alpha* coefficient and yielded a value of 0,744. Based on the criteria proposed by this value falls within the range of $0{,}60 \leq \alpha < 0.80$ and is therefore categorized as having high reliability, indicating good internal consistency of the instrument. This reliability coefficient is higher than that reported by Basri & Turmuzi, (2021) this value falls within the range of $0{,}60 \leq \alpha < 0{,}80$ and is therefore categorized as having high reliability, indicating good internal consistency of the instrument. This reliability coefficient is higher than that reported by Hapsari, (2017) who obtained a reliability coefficient of 0.68 in the development of a similar mathematical literacy test. Therefore, the instrument developed in the present study demonstrates superior measurement quality.

The analysis of item difficulty was conducted according to the criteria proposed by Pakpahan (1990) as presented in Table 4. The results showed that 2 items (13.3%) were classified as difficult, with difficulty indices ranging from 0,10 to 0,29, namely Item 13 (p = 0,25) and Item 15 (p = 0,29). Meanwhile, the remaining 13 items (86.7%) were categorized as moderate, with difficulty indices ranging from 0,30 to 0,70. According to Pakpahan (1990) items within this range are considered acceptable and suitable for use. The predominance of moderately difficult items indicates that the instrument is appropriate for assessing students across a range of ability levels. Item discrimination was analyzed based on the criteria established by Crocker & Algina, (1986) as presented in Table 5. The results indicated that 4 items (26.7%) demonstrated good discrimination, with values ranging from 0,40 to 1,00, namely Items 2, 4, 9, and 12. In addition, 10 items (66.7%) had discrimination indices ranging from 0,30 to 0,39 and were therefore categorized as acceptable without revision, namely Items 1, 3, 5, 6, 7, 8, 10, 11, and 14. Item 13, with a discrimination index of 0,25, fell within the range of 0,20 – 0,29 and was therefore categorized as requiring revision. Based on these criteria, no items were rejected because none had discrimination indices below 0,19. Consequently, all 15 items were considered suitable for progression to the *field-test* stage, although Item 13 required revision to improve its discrimination power and enhance the overall quality of the instrument.

The *field test* was conducted in four schools: SMP Negeri 1 Banda Aceh (Class IX-8), SMP Negeri 3 Banda Aceh (Class IX-8), SMP Negeri 17 Banda Aceh (Class IX-8), and MTsN 1 Banda Aceh (Class IX-10), involving a total of 130 students. The schools were selected based on differences in accreditation status and student characteristics to ensure representation of the target population. The purpose of this stage was to obtain empirical data on item quality on a larger scale, allowing validity, reliability, item difficulty, discrimination power, and distractor effectiveness to be analyzed more accurately. In addition, 36 students were selected to complete a response questionnaire to obtain their perceptions of the developed test items. The results of the analyses of validity, reliability, item difficulty, discrimination power, and distractor effectiveness are presented in Table 10.

Tabel 10. Rekapitulation of item analysis in the *field test stage* (N=130)

Item no.	Proportion correct (P)	Category	*Biserial*	*Point Biserial*	Remarks	Distractor function
1	0,769	Easy	0,535	0,386	Valid	A, B ≥ 5%; D ≤ 5%
2	0,646	Moderate	0,760	0,591	Valid	A, C, D ≥ 5%
3	0,631	Moderate	0,841	0,658	Valid	A, B, D ≥ 5%
4	0,754	Easy	0,627	0,459	Valid	A, D ≥ 5%; B ≤ 5%

5	0,369	Moderate	0,437	0,342	Valid	A, C, D ≥ 5%
6	0,662	Moderate	0,677	0,523	Valid	A,C ≥ 5%; D ≤ 5%
7	0,585	Moderate	0,577	0,457	Valid	A, B, D ≥ 5%
8	0,585	Moderate	0,629	0,497	Valid	A,C ≥ 5%; D ≤ 5%
9	0,315	Moderate	0,552	0,423	Valid	A, B, D ≥ 5%
10	0,562	Moderate	0,620	0,492	Valid	A, B, D ≥ 5%
11	0,477	Moderate	0,606	0,484	Valid	A, C, D ≥ 5%
12	0,415	Moderate	0,478	0,378	Valid	A, C, D ≥ 5%
13	0,300	Sedang	0,449	0,340	Valid	A, B, D ≥ 5%
14	0,331	Moderate	0,696	0,536	Valid	B, C, D ≥ 5%
15	0,485	Moderate	0,895	0,714	Valid	B, C, D ≥ 5%

Based on the results of the analysis conducted using the ITEMAN program on data from 130 students, all 15 test items were declared valid, with *point-biserial* coefficients ranging from 0.340 to 0.714. According to Crocker & Algina (1986) items with *point-biserial* coefficients above 0.30 possess adequate validity. Therefore, all items in the instrument satisfied the criteria for empirical validity. Furthermore, the biserial coefficients, which ranged from 0.437 to 0.895, indicate a strong correlation between individual item scores and the total test score (Allen & Yen, 1979).

The test reliability coefficient obtained was 0.770. Based on the criteria proposed by Basri & Turmuzi, (2021), this value falls within the range of $0.60 \leq r_{11} < 0.80$ and is therefore categorized as having high reliability. This finding indicates that the instrument has good internal consistency and a relatively small measurement error (*Standard Error of Measurement* = 1.668). The reliability coefficient obtained in this study is comparable to that reported by Wulandari *et al.* (2020) who obtained a reliability coefficient of 0.78 in the development of Higher Order Thinking Skills (HOTS) items on algebraic content.

The analysis of item difficulty was conducted based on the criteria proposed by Pakpahan (1990), as presented in Table 4. The results showed that 2 items (13.3%) were classified as easy, with difficulty indices ranging from 0.71 to 0.90, namely Item 1 (p = 0.769) and Item 4 (p = 0.754). According to the established criteria, these items were categorized as good items but recommended for revision. In addition, 12 items (80.0%) were classified as having moderate difficulty, with difficulty indices ranging from 0.30 to 0.70, indicating that they were good items and could be used without revision. Although the distribution of item difficulty in this study was dominated by moderately difficult items, this condition remains acceptable because the instrument was designed to assess the mathematical literacy abilities of Grade IX students at an intermediate level of proficiencys.

Item discrimination was analyzed according to the criteria proposed by Crocker & Algina (1986) as presented in Table 5. The results indicated that all 15 items had biserial coefficients above 0.40 and were therefore classified as good (0.40–1.00). This finding demonstrates that each item was able to effectively distinguish between students with high ability and those with low ability (Prayekti, 2024). No items fell into the categories of acceptable without revision, requiring revision, or rejected.

The analysis of distractor effectiveness was based on the criterion that a distractor is considered effective if it is selected by at least 5% of test participants (*Analisis Butir Soal Dengan BIGSTEPS*, 1997; Arhin, 2024). The results showed that 9 items (60%) had three distractors functioning effectively and were therefore categorized as very good, namely Items 2, 3, 5, 9, 10, 11, 12, 13, and 15. Meanwhile, 6 items (40%) had two effective distractors and were categorized as good, namely Items 1, 4, 6, 7, 8, and 14. No items were found to have only one functioning distractor or no functioning distractors at all. Overall, the distractor effectiveness of the instrument was categorized as very good because every item contained at least two distractors that functioned effectively.

Nevertheless, several distractors did not function optimally because they were selected by fewer than 5% of the students. These included option B in Item 4, option D in Items 6 and 8, and option C in Item 14. This condition may have occurred because the numerical values presented in these distractors were too far from a plausible answer, or because the distractors were not fully constructed based on students' common misconceptions in geometry. As a result, even students with lower ability levels were not inclined to select these options. To improve distractor quality, redesign efforts should be based on a more in-depth analysis of student errors, such as interviews or literature reviews focusing on common misconceptions in geometry. These findings indicate that most distractors developed based on students' typical error patterns (Duskri *et al.*, 2014), successfully attracted students who had not yet fully mastered the concepts, thereby fulfilling their diagnostic function. This result is consistent with (Retnawati, 2016) who stated that effective distractors are selected by more than 5% of test participants and tend to be chosen more frequently by low-achieving students. To further demonstrate that the developed instrument not only satisfies psychometric criteria but also substantively measures mathematical literacy, an analysis of one representative test item is presented below. Item 15 was selected as an example because it achieved the highest point-biserial coefficient (0.714) and a biserial coefficient of 0.895. In addition, all of its distractors functioned effectively, with options B, C, and D each being selected by at least 5% of the participants.

<u>Wacana V</u>

Bacalah wacana berikut untuk menjawab soal nomor 14-15.

Desain Atap Gudang

Pak Andi adalah seorang arsitek yang sedang merancang sebuah gudang penyimpanan logistik. Gudang ini akan digunakan untuk menyimpan berbagai jenis barang logistik seperti bahan makanan pokok (beras, gula, minyak), produk kemasan, alat tulis kantor, serta peralatan usaha lainnya. Penyimpanan barang-barang tersebut memerlukan ruangan dengan sirkulasi udara yang baik agar tidak lembap dan menjaga kualitas barang.

Untuk memenuhi kebutuhan sirkulasi udara tersebut, Pak Andi mendesain atap gudang berbentuk segitiga sama kaki yang menonjol di bagian depan. Desain ini tidak hanya berfungsi untuk estetika, tetapi juga memungkinkan udara panas naik dan keluar melalui ventilasi di bagian atas atap, sehingga suhu di dalam gudang tetap terjaga.

Selain struktur atap, Pak Andi juga harus memperhitungkan penempatan tangki cadangan air berbentuk tabung di dalam gudang. Tangki ini merupakan bagian dari sistem pemadam api mandiri yang sangat penting mengingat barang logistik yang disimpan banyak yang termasuk kategori mudah terbakar seperti kertas, kemasan kardus, dan bahan lainnya. Volume tangki harus cukup untuk menanggulangi kemungkinan kebakaran tanpa mengambil terlalu banyak ruang penyimpanan. Efisiensi dalam penggunaan bahan atap dan penentuan volume tangki menjadi kunci utama dalam perancangan ini, agar pembangunan gudang tetap ekonomis namun tetap fungsional dan aman sesuai standar penyimpanan logistik.

Figure 2. Visualization of a Logistics Warehouse Design

15. Di dalam gudang tersebut, Pak Andi menempatkan tangki air berbentuk tabung dengan diameter 2 meter dan tinggi 3 meter. Untuk biaya operasional, pemilik gudang menetapkan anggaran Rp5.000 per meter kubik air. Berdasarkan spesifikasi tersebut, pernyataan yang paling tepat untuk mengevaluasi kebutuhan biaya pengisian penuh tangki tersebut adalah... *(Gunakan π = 3,14)*

A. Volume tangki sekitar 9,42 m^3, sehingga biayanya adalah Rp47.100.
B. Volume tangki sekitar 18,84 m^3, sehingga biayanya mencapai Rp94.200.
C. Luas alas tangki adalah 6,28 m^3, sehingga biaya ditentukan oleh luas permukaan, bukan volume.
D. Karena diameter 2 meter, maka volume tangki adalah 37,68 m^3, dan membutuhkan biaya lebih dari Rp150.000.

The item above satisfies all four mathematical literacy indicators proposed by Ayuningtyas (2017) which served as the framework for this study. *First*, students identify mathematical information from a real-life context, namely a diameter of 2 m, a height of 3 m, and a cost of IDR 5,000 per cubic meter. *Second*, students transform this information into a mathematical model by converting the diameter into a radius (r = 1 m) and then applying the cylinder volume formula, $V = \pi \times r^2 \times t$. *Third*, students use the model to calculate the volume, $V = 3.14 \times 1^2 \times 3 = 9.42$ m^3, and subsequently determine the cost as 9.42 × IDR 5,000 = IDR 47,100. *Fourth*, students evaluate the reasonableness of all four statements (A–D) both mathematically and contextually, rather than merely calculating a single correct answer. Distractors B, C, and D were designed based on common student errors, including using the diameter instead of the radius, confusing area with volume, and making computational errors. Therefore, this item assesses not only procedural skills but also students' ability to formulate, apply, and evaluate mathematical solutions within an occupational context, as conceptualized by OECD (2023) Consequently, the item aligns with the mathematical literacy processes of formulating, employing, and interpreting and evaluating.

Descriptive statistics of students' total scores are presented in Table 11.

Tabel 11. Descriptive statistics for students' total scores

Statistic	Nilai
Number of students	130
Mean	7,885
Variance	12,087
Standard deviation	3,477

Skewness	-0,043
Kurtosis	-1,395
Score Minimum	2,000
Score maksimum	14,000
Median	8,000

The descriptive statistics of students' total scores indicate a mean score of 7.885 out of a maximum possible score of 15, with a standard deviation of 3.477. The relatively large standard deviation suggests considerable variation in students' abilities (Freedman *et al.*, 2007; (Ravid, 2020). The score range, extending from 2 to 14, indicates the absence of extreme values that might distort the analysis results. Furthermore, the skewness value of −0.043 and the kurtosis value of −1.395 both fall within the acceptable range of −2 to +2, indicating that the data can be considered normally distributed (Ghozali, 2016; Field, 2018; George & Mallery, 2019). Therefore, the developed mathematical literacy test instrument meets the criteria of a valid and reliable tool for measuring the mathematical literacy abilities of Grade IX junior secondary school students.

Consistent with the quantitative findings indicating that the quality of the test items was generally good, the questionnaire data obtained from 36 students also supported these results. A total of 83% of students provided positive responses regarding the usefulness of the test items, placing this aspect in the *very good* category. This finding suggests that the developed items helped students understand geometry concepts and solve contextual problems. The attractiveness aspect received a percentage score of 69%, which falls into the *good* category, while the difficulty level aspect obtained a percentage score of 55%, categorized as *moderate*. The positive responses, particularly regarding usefulness and attractiveness, indicate that the integration of individual and occupational contexts into the mathematical literacy items successfully achieved its intended purpose of making geometry learning more meaningful and relevant to students' lives. These findings support the study by Sari, R., & Putra (2021), which reported that context-based mathematics problems can enhance students' motivation and participation in mathematics learning.

Table 11. Summary of student questionnaire results by aspect

No	Aspect	Presentase	Kategori
1	Item attractiveness	69 %	Good
2	Difficulty level	55 %	Moderate
3	Item userfulness	83 %	Very good

This study has several limitations that should be acknowledged. First, the study focused exclusively on the development of mathematical literacy items in the geometry content domain using individual and occupational contexts for Grade IX junior secondary school students. Therefore, the findings cannot yet be generalized to other mathematical domains, such as algebra, numbers, data, and probability, or to different grade levels. Second, the developed instrument consisted solely of multiple-choice items and did not include open-ended questions that could provide deeper insights into students' thinking processes, including their problem-solving strategies and mathematical reasoning abilities. Third, due to limitations in time, resources, and funding, the development process was conducted only up to the *field-test* stage, involving 130 students from four schools in Banda Aceh. Consequently, the findings should be generalized cautiously when applied to broader populations with different cultural and socioeconomic backgrounds. Fourth, the *one-to-one* stage involved only six students, which limited the study's ability to generate statistically robust validity evidence, although substantively it provided valuable input for item improvement. Fifth, several distractors in certain items did not function optimally because they were selected by fewer than 5% of students. Therefore, further refinement is necessary to ensure that all distractors function effectively.

Conclusion

This study aimed to develop a mathematical literacy test on the geometry content domain using individual and *occupational* contexts for Grade IX junior secondary school students through Tessmer's development model, which consists of the *preliminary*, *self-evaluation*, *prototyping* (*expert review*, *one-to-one*, and *small-group evaluation*), and *field-test* stages. During the *preliminary* stage, curriculum, student, and context analyses were conducted, leading to the decision to integrate individual and occupational contexts into 15 multiple-choice items covering cognitive levels C2 to C5.

During the *expert review* stage, all test items were judged to be highly valid by three validators, with an average validation score of 87% (very valid category), comprising content validity (93%), construction validity (86%), context validity (90%), and language validity (80%). Based on the validators' feedback, several editorial and technical revisions were made to *Prototype I*, resulting in *Prototype II*. During the *one-to-one* stage involving six students, further revisions were implemented to improve items that were considered unclear in terms of language and illustrations. *Prototype II* was subsequently tested in the *small-group evaluation* stage involving 31 students, where all items were found to be valid with a reliability coefficient of 0.744 (high reliability category). The item difficulty analysis indicated that 13 items were classified as moderate and 2 items as difficult. The discrimination analysis showed that 4 items were categorized as good (0.40–1.00), 10 items were accepted without revision (0.30–0.39), and 1 item required revision (0.20–0.29), with no items being rejected.

During the *field-test* stage involving 130 students from four schools in Banda Aceh, all 15 items were declared valid, with *point-biserial* coefficients ranging from 0.340 to 0.714 and a reliability coefficient of 0.770 (high reliability category). The item difficulty distribution consisted of 2 easy items requiring revision and 13 moderate items suitable for use. All items demonstrated good discrimination power, with biserial coefficients exceeding 0.40. The distractor effectiveness analysis showed that 9 items (60%) were categorized as very good and 6 items (40%) as good. Nevertheless, four distractors—option B in Item 4, option D in Items 6 and 8, and option C in Item 14—require revision because they were selected by fewer than 5% of students. Student responses were also positive, with 83% indicating that the test items were beneficial for learning.

Considering the findings and limitations of this study, including its focus on geometry content, the exclusive use of multiple-choice items, the limited sample drawn from Banda Aceh, and the presence of several suboptimal distractors, future research is recommended to develop similar instruments in other mathematical domains, incorporate open-ended items, involve larger and more geographically diverse samples, refine distractors that do not function optimally, and analyze item characteristics using *Item Response Theory* (IRT). Furthermore, future studies should investigate the effectiveness of these items in improving students' mathematical literacy through experimental research designs and explore the integration of context-based mathematical literacy tasks into classroom instruction.

Acknowledgment

The author would like to express sincere gratitude to Dr. M. Duskri, M.Kes., as the research supervisor, for his invaluable guidance, support, and encouragement throughout the conduct of this study. The author also extends heartfelt appreciation to the expert validators, namely Kamarullah, M.Pd. (Educational Evaluation Lecturer), Rahmi Fuadi, S.Pd.I., M.Pd., and Cut Anggraini, S.Pd.Gr. (Junior Secondary School Mathematics Teachers), for their valuable suggestions, insightful comments, and constructive feedback that contributed significantly to the improvement of the research instrument. The author further wishes to thank the principals, teachers, and students of SMP Negeri 1 Banda Aceh, SMP Negeri 3 Banda Aceh, SMP Negeri 17 Banda Aceh, and MTsN 1 Banda Aceh for their cooperation, participation, and support throughout the implementation of this study.

References

Adinawan, C. (2016). *Matematika untuk SMP/MTs kelas IX*. Erlangga.

Aldiansyah, R. W., & Agoestanto, A. (2024). Analisis kemampuan literasi matematika ditinjau dari kemandirian belajar siswa SMP. *PRISMA Prosiding Seminar Nasional Matematika*, *7*, 123–130.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Analisis Butir Soal dengan BIGSTEPS. (1997). Pulitbang Sisjian Depdikbud.

Arhin, A. K. (2024). Developing Distractors for Mathematics Multiple Choice Items: A Literature Review. *Acta Educationis Generalis*, *14*(3), 103–120. https://doi.org/10.2478/atd-2024-0022

Ayuningtyas, N. (2017). Profil Literasi Matematis Konten Change and Relationship Siswa Kelas X Ditinjau dari Gaya Kognitif Visualizer dan Verbalizer. *Jurnal Edukasi*, *3*(1), 99–109.

Ayutias, P., Husna, N., & Buyung. (2024). Profil kemampuan literasi matematis siswa kelas VII di SMP Negeri 9 Singkawang. *Pentagon: Jurnal Matematika Dan Ilmu Pengetahuan Alam*, *2*(3), 15–22. https://doi.org/10.62383/pentagon.v2i3.195

Bano, V. O., Marambaawang, D. N., & Njoeroemana, Y. (2022). Analisis kriteria butir soal ujian sekolah mata pelajaran IPA di SMP Negeri 1 Waingapu. *Ideas: Jurnal Pendidikan, Sosial, Dan Budaya*, *8*(1), 145–152. https://doi.org/10.32884/ideas.v8i1.660

Basri, K., & Turmuzi, M. (2021). Analisis Butir Soal Ulangan Semester Ganjil Mata Pelajaran Matematika Kelas VIII SMP Pada Tahun Ajaran 2018/2019. *Griya Journal Of Mathematics Education and Aplication*, *1*(4), 682–694.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Duskri, M., Kumaidi, & Suryanto. (2014). Developing diagnostic test of mathematics learning difficulties in elementary schools. *Laporan Penelitian. Universitas Negeri Yogyakarta.*

Fadilah, N., & Sutrisno, S. (2021). Pengembangan instrumen penilaian HOTS berbasis model Tessmer pada materi geometri SMP. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *5*(2), 1458–1473.

Fauzi, A., & Arifin, Z. (2020). Pembelajaran geometri berbasis konteks: Studi kasus di sekolah menengah pertama. *Jurnal Pendidikan Matematika*, *15*(2), 123–134.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education (8th ed.)*. McGraw-Hill.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics (4th ed.)*. W. W. Norton & Company.

Ghozali, I. (2016). *Aplikasi analisis multivariate dengan program IBM SPSS*. Badan Penerbit Universitas Diponegoro.
Grounlund, N. E. (1982). *Counstructing Achievement Tesis*. Prentice-Hall, Inc.
Hapsari, T. (2017). Literasi matematis siswa. *Jurnal Euclid*, *6*(1).
Isnaniah, & Imamuddin, M. (2022). Pengembangan soal literasi matematika konteks budaya Minangkabau untuk meningkatkan literasi matematika siswa. *Jurnal UIN Bukit Tinggi*, *11*(5), 3716–3726.
Kemendikbud. (2019). *Laporan hasil ujian nasional (UN) SMP/MTs tahun 2019*. Kementerian Pendidikan dan Kebudayaan.
Kemendikbud. (2020). *Panduan Asesmen Kompetensi Minimum (AKM)*. Kementerian Pendidikan dan Kebudayaan.
Kemendikbudristek. (2025). *Laporan hasil Asesmen Nasional tahun 2024*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.
Masfufah, & Afriansyah, E. A. (2021). Analisis kemampuan literasi matematis siswa melalui soal PISA. *Jurnal Pendidikan Matematika*, *15*(1), 1–12.
Masrun. (1979). *Pengukuran Pendidikan*. Fak. Psikologi UGM.
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. TIMSS & PIRLS International Study Center.
Mulyatiningsih, E. (2011). *Metode penelitian terapan bidang pendidikan*. Alfabeta.
NCTM. (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.
OECD. (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing.
OECD. (2023). *PISA 2022 results (Volume I and II)*. OECD Publishing.
Pakpahan, P. (1990). *Analisi Soal Berdasarkan Data Empirik*. Pusisjian Depdikbud.
Prayekti, N. (2024). *Evaluasi pembelajaran matematika*. Literasi Nusantara.
Qadir, A., Huda, N., & Hermmina, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda dan Efektifitas Pengecoh. *Al Furqan: Jurnal Agam, Sosial,Dan Budaya*, *3*(3), 1450–1467.
Rahayu, S. (2020). Penerapan teori perkembangan kognitif Piaget dalam pembelajaran matematika SMP. *Jurnal Pendidikan Matematika, 15*(1), 45–60.
Rahmayani, R., Sukayasa, Ismaimuza, D., & Meinarni, W. (2024). Analysis of mathematical literacy skills of students in class VIII SMP Negeri 3 Dampelas in solving geometry problems in terms of Van Hiele level. *Prima: Jurnal Pendidikan Matematika*, *8*(2), 282–292. https://doi.org/10.12345/prima.v8i2.10888
Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Parama Publishing.
Santoso, A. (2017). Benarkah ukuran sampel minimal = 30? *Jurnal Psikologi Indonesia*, *12*, 63–84.
Sari, R., & Putra, D. (2021). Pengembangan soal literasi matematis berbasis konteks untuk siswa SMP. *Jurnal Ilmiah Pendidikan Matematika*, *10*(1), 67–78.
Sari, R., & Putra, D. (2021). Pengembangan soal literasi matematis berbasis konteks untuk siswa SMP. *Jurnal Ilmiah Pendidikan Matematika*, *10*(1), 67–78.
Suryabrata, S. (1987). *Pengembangan Tes Hasil Belajar*. Rajawali Press.
Syafni, Z. Z. (2017). *Dasar-dasar Ilmu Pendidikan*. Kencana.
Tessmer, M. (1993). *Planning and conducting formative evaluations: Improving the quality of education and training*. Kogan Page.
Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey-Bass.
Utami, M. R. ., Hapizah, & Mulyono, B. (2025). Pengembangan LKPD Digital Berbantuan GeoGebra Materi Persamaan Linear Satu Variabel untuk Mendukung Kemampuan Computational Thinking. *Media Pendidikan Matematika*, *13*(2), 865–884.
Widyawati, H., & Duskri, M. (2025). Development of Algebraic Diagnostic Tests for Grade VII Students of State Islamic Junior High School. *MaPan: Jurnal Matematika Dan Pembelajaran*, *13*(2), 383–404.

Wijaya, A. (2018). Students' difficulties in understanding geometry concepts: A qualitative study. *Journal of Educational Research*, *12*(3), 45–56.
Wilson, L. O. (2016). *Anderson and Krathwohl Bloom's taxonomy revised*.
Wulandari, S., Hajidin, H., & Duskri, M. (2020). Pengembangan soal higher order thinking skills (HOTS) pada materi aljabar di sekolah menengah pertama. *Jurnal Didaktik Matematika, 7*(2), 200–220.