MaPan : Jurnal Matematika dan Pembelajaran
p-ISSN: 2354-6883 ; e-ISSN: 2581-172X
Volume 13, No 2, Dec 2025 (383-404)
DOI: https://doi.org/10.24252/mapan.2025v13n2a8

Development of Algebraic Diagnostic Tests for Grade VII Students of State Islamic Junior High School

Hesti Widyawati[1], M. Duskri[2*]
[1,2]Department of Mathematics Education, UIN Ar-Raniry Banda Aceh
[1,2]Jl. Syeikh Abdul Rauf Kopelma Darussalam Banda Aceh, Indonesia
Correspondence Email: m.duskri@ar-raniry.ac.id

Received October 7, 2025; Revised December 18, 2025; Accepted December 23, 2025
Available Online December 29, 2025

Abstract:
Difficulties in learning algebra are a common problem experienced by grade VII students, especially in understanding the symbols, variables, and basic concepts of algebra. These difficulties not only have an impact on low understanding of concepts, but also have the potential to lead to misconceptions that hinder advanced mathematics learning, such as equations, functions, and advanced algebraic materials. Therefore, diagnostic instruments are needed that are able to accurately identify students' learning difficulties and misconceptions from an early age. This study aims to develop a valid and reliable two level multiple-choice diagnostic test instrument to identify students' learning difficulties and misconceptions in algebraic material. The research uses the Research and Development (R&D) method with the Tessmer model, which includes preliminary stages, self-evaluation, expert review, one to one, small group, and field trials. The research subjects consisted of 60 students in grades VII-7 and VII-10 MTsN 1 Banda Aceh. The analysis of question items was carried out quantitatively using Iteman software to assess the validity, reliability, level of difficulty, differentiation, and effectiveness of the distractor. The results of the development showed that of the 38 questions developed, 22 were accepted, 10 revised, and 6 were discarded, with a reliability coefficient of 0.836, which is classified as very high. The resulting diagnostic tests are effective in identifying learning difficulties and student misconceptions, so that they have a scientific contribution as a quality diagnostic evaluation instrument and practical benefits for teachers in designing learning and remedial programs that are more targeted.

Abstrak:
Kesulitan belajar aljabar merupakan permasalahan umum yang dialami oleh peserta didik kelas VII, khususnya dalam memahami simbol, variabel, dan konsep dasar aljabar. Kesulitan ini tidak hanya berdampak pada rendahnya pemahaman konsep, tetapi juga berpotensi menimbulkan miskonsepsi yang dapat menghambat pembelajaran matematika lanjutan, seperti persamaan, fungsi, dan materi aljabar tingkat lanjut, sehingga diperlukan instrumen diagnostik yang mampu mengidentifikasi secara akurat kesulitan belajar dan miskonsepsi peserta didik sejak dini. Penelitian ini bertujuan untuk mengembangkan instrumen tes diagnostik pilihan ganda dua tingkat yang valid dan reliabel guna mengidentifikasi kesulitan belajar dan miskonsepsi peserta didik pada materi aljabar. Penelitian ini menggunakan

metode Research and Development (R&D) dengan model Tessmer yang meliputi tahap pendahuluan (preliminary), evaluasi diri (self evaluation), telaah ahli (expert review), uji coba satu-satu (one to one), kelompok kecil (small group), dan uji lapangan (field test). Subjek penelitian terdiri atas 60 peserta didik kelas VII-7 dan VII-10 di MTsN 1 Banda Aceh. Analisis butir soal dilakukan secara kuantitatif menggunakan perangkat lunak Iteman untuk menilai validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Hasil pengembangan menunjukkan bahwa dari 38 butir soal yang dikembangkan, sebanyak 22 butir diterima, 10 butir direvisi, dan 6 butir dibuang, dengan koefisien reliabilitas sebesar 0,836 yang tergolong sangat tinggi. Tes diagnostik yang dihasilkan efektif dalam mengidentifikasi kesulitan belajar dan miskonsepsi peserta didik, sehingga memberikan kontribusi ilmiah sebagai instrumen evaluasi diagnostik yang berkualitas serta manfaat praktis bagi guru dalam merancang pembelajaran dan program remedial yang lebih tepat sasaran.

Keywords:
Algebra, Diagnostic Tests, Learning Difficulties, Two-Level Multiple Choice, Tessmer Model

How to Cite: Duskri, M., & Widyawati, H. (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. https://doi.org/10.24252/mapan.2025v13n2a8.

INTRODUCTION

Mathematics education has a strategic role in developing students' logical, critical, and creative thinking skills. As a subject taught at all levels of education, mathematics not only functions as a collection of concepts and procedures but also as a means to practice problem-solving skills needed in daily life (Nabilah, Amalia, Angreini, Rahmi, Zulkarnain, & Fajriah, 2024). Therefore, mathematics learning needs to be designed systematically to be able to develop students' potential optimally, both from cognitive, affective, and psychomotor aspects.

One of the branches of mathematics that has a fundamental role in the development of abstract thinking skills is algebra (Sinabang, Lumbantoruan, Morina, Sagala, & Tanjung, 2025). Algebra material began to be formally introduced at the junior high school level as a continuation of the concept of arithmetic in elementary school (Rahayu, Kurniati, Jatmiko, Lestari, & Ambarwati, 2022). Algebra learning requires students to understand algebraic symbols, variables, and operations as the basis for learning advanced math

material, such as equations, functions, and analytical geometry. Good mastery of algebra has been proven to contribute to the academic success of students at the next level of education, especially in the fields of mathematics and science (Sinabang, Lumbantoruan, Morina, Sagala, & Tanjung, 2025). Algebraic material also serves to introduce symbolization, variables, and manipulation of equations, which are the basis for the material on functions, analytical geometry, and science/engineering lessons at the next level. Algebra provides a solid foundation for understanding more complex mathematical concepts, such as functions and calculus. Students who master algebra will be better prepared for advanced math material in high school and college. Research shows that understanding algebra contributes to future academic success in math and science (McEachin, Domina, & Penner, 2020). In addition, learning algebra encourages the ability to generalize patterns and abstract reasoning, which is important in various disciplines (Dahiana, Herman, Nurlaelah, & Pereira, 2023).

However, learning algebra often faces significant challenges. Many students struggle to understand algebraic symbols and manipulation rules, resulting in poor conceptual understanding. This difficulty can lead to misunderstandings, which include a misunderstanding of mathematical concepts. Research Nugraha, Kadarisma, and Setiawan (2019) found that students often make mistakes, revealing that student mistakes in algebra generally include errors in number operations, errors in understanding problems, procedural errors, and calculation errors. In addition, the phenomenon of guessing answers and beliefs about wrong answers shows the weak conceptual understanding of students. This condition has the potential to hinder the continuous mathematics learning process if it is not identified early. In addition to misunderstandings, the phenomenon of guessing and guessing answers is also common. Misunderstandings occur when students have high confidence in the wrong answers. Guessing occurs when students choose the correct answer but are unable to give a proper reason and are unsure of their choice. Guess luck refers to a situation in which a student chooses the correct answer with valid reasons, but without complete confidence. These three conditions reflect a weak conceptual understanding that has the potential to hinder optimal mathematics learning.

These learning difficulties and misconceptions have a direct impact on students' low learning outcomes, including the inability to meet the Minimum Completeness Criteria (Widayanti, Rahayuningsih, Yuliana, & Christian, 2025).

Therefore, systematic efforts are needed to accurately detect students' learning difficulties. One relevant approach is the use of diagnostic tests. Diagnostic tests aim to identify specific learning difficulties and misconceptions of students so that teachers can design learning and remedial programs that are more targeted (Duskri, Kumaidi, & Suryanto, 2014; Sriyanti, Mania, & Hairani, 2019). In the implementation of the Independent Curriculum, diagnostic tests play a strategic role as the first step in mapping students' abilities and identifying their learning needs (Triyono, Masrukan, & Mulyono, 2023). With this information, teachers can develop a more targeted and effective learning plan. Diagnostic tests vary in format, from one level, two levels, three levels, to four levels.

Theoretically, the development of diagnostic tests is based on the theory of mathematical learning difficulties and misconception theory, which emphasizes the importance of revealing the structure of students' understanding, not just the right or wrong answers (Nursalam, 2016). One form of diagnostic test that is considered effective is the two tier multiple choice test, which combines multiple choice answers with the reason for the selection of the answer. This format allows educators to identify conceptual errors and student thinking patterns in greater depth than a single-level test, and is relatively appropriate to the characteristics of junior high school students.

Various previous studies have developed mathematical diagnostic tests, but most of them still focus on the form of description or have not specifically examined the quality of the question items based on comprehensive empirical analysis. In addition, the development of algebraic diagnostic tests in the form of two level multiple choice tests for grade VII junior high school students with quantitative analysis of the characteristics of question items is still limited. This gap is the basis for the development of diagnostic test instruments that are valid, reliable, and able to accurately reveal learning difficulties and misconceptions of students.

Based on the results of initial observations and interviews with grade VII mathematics teachers at MTsN 1 Banda Aceh, information was obtained that some students still had difficulties in understanding basic algebraic concepts. Difficulties that often arise include the use of algebraic symbols, the determination of variable values, and the application of positive and negative number operations in algebraic form. The teacher also said that students tend to make repeated mistakes even though the material has been delivered, and

the assessment used so far has not been able to fully reveal the source of the difficulty or the reason behind the students' answers. These empirical findings show the need for an evaluation instrument that can identify students' learning difficulties in more depth, especially in algebraic materials.

Based on this description, this study focuses on the development of a two-level multiple-choice diagnostic test on algebra material for seventh-grade junior high school students. This research is expected to contribute scientifically to the development of valid and reliable mathematical diagnostic assessment instruments and practically to assist teachers in mapping students' learning difficulties and designing more effective and targeted learning strategies.

The developed diagnostic test is designed based on the concept map and subtopics of seventh-grade algebra to ensure continuity of the assessed material. The alternative answers, particularly the distractors, are constructed based on common student error patterns, including conceptual errors, procedural errors, carelessness, misunderstandings of positive and negative number operations, and misinterpretation of problem statements. In addition, students are required to provide reasons for their selected answers, enabling a more accurate identification of students' learning difficulties and misconceptions. Through this approach, each selected distractor can be used to map specific learning difficulties, thereby supporting teachers in planning appropriate remedial instruction.

Previous studies have shown that diagnostic tests are effective in identifying students' difficulties in learning mathematics, including algebra. However, many existing instruments focus primarily on students' final answers and do not adequately reveal the underlying reasoning or thinking patterns. Therefore, this study develops a two-level multiple-choice diagnostic test tailored to the characteristics of seventh-grade algebra material. The instrument development follows the formative evaluation stages of the Tessmer model to ensure content relevance, language clarity, and item quality. Consequently, the resulting instrument is expected to provide a more comprehensive description of students' algebra learning difficulties and serve as a useful consideration for teachers in designing instruction that aligns with students' learning needs.

METHODS

This research uses the Research and Development (R&D) method with reference to the Tessmer model. The selection of the Tessmer model is based on its characteristics that emphasize gradual formative evaluation, making it ideal for the development of diagnostic test instruments. This model allows researchers to make repeated revisions based on expert input and learner responses, so that the resulting product has a high level of validity and reliability. In addition, the Tessmer model is widely used in the development of learning evaluation instruments because of its systematic and easily replicated procedures (Sutiani, Herawati, & Hermanto, 2024; Wulandari, Hajidin, & Duskri, 2020). The research was carried out at MTsN 1 Banda Aceh with 60 research subjects in grades VII-7 and VII-10. The sampling technique used is purposive sampling, with the consideration that both classes have studied algebraic material in accordance with the learning outcomes of grade VII and have heterogeneous academic characteristics. The selection of this sample aims to obtain a representative picture of students' learning difficulties in algebraic materials.

Methodologically, the testing of the instrument at the small group stage with an adequate sample size has provided a preliminary picture of the characteristics of the question items, including the level of difficulty, differentiation, and effectiveness of the deception, which are then analyzed based on the quality criteria of the item that has been determined. The limitation of testing at this stage is carried out because the main purpose of the research focuses on the development and testing of the initial quality of diagnostic test instruments, especially on the aspects of content validity, construction clarity, and question item characteristics. The small group stage is considered sufficient to obtain the necessary information as a basis for refining the instrument before it is applied on a wider scale. Consideration of limited time and access to research subjects is also the reason for not carrying out large-scale field tests. Nonetheless, the number of samples in the small group trial was eligible for item analysis, so the information obtained remained representative and could be used as a basis for decision-making on the quality of each item (Singarimbun & Effendi, 1995).

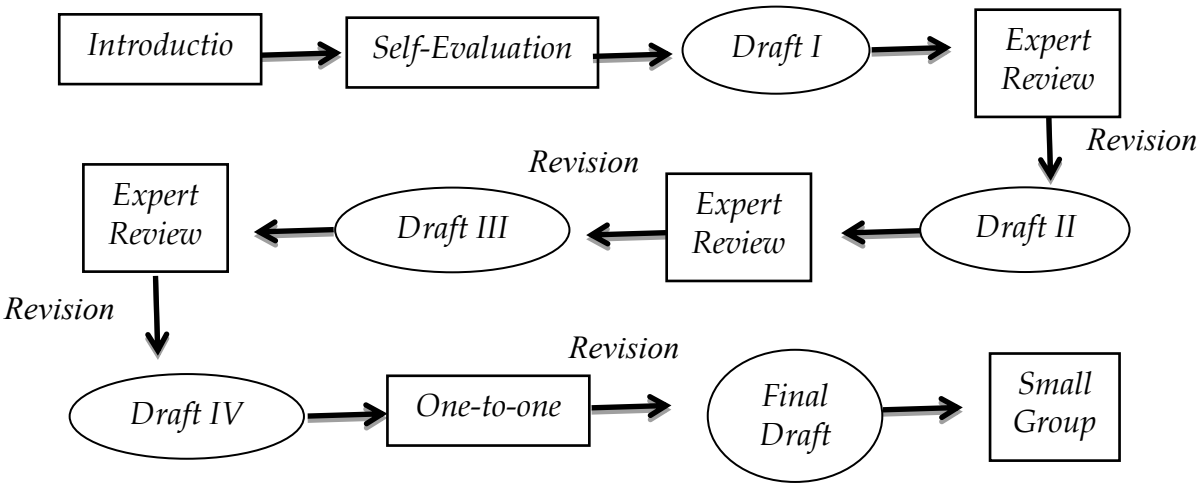

Figure 1. Working Procedure

The procedure for developing diagnostic test instruments in this study is shown in figure 1. The image depicts the stages of development based on the Tessmer model, which includes the preliminary stage, self evaluation, expert review, one to one test, and small group test. Each stage is carried out sequentially to ensure that the instrument developed is of adequate quality before being used at the test stage.

The development of diagnostic test instruments is carried out through the stages of the Tessmer model, which include the initial stage (preliminary), self evaluation, expert review, one to one, and small group (small group). In this study, the development process was limited to the small group stage. The initial stage consists of two main activities, namely preparation and design. In the preparation stage, an analysis of the curriculum, teaching materials, and student characteristics was carried out through document studies and interviews with grade VII mathematics teachers. The design stage is focused on the preparation of a question grid and the development of a two level multiple choice diagnostic test, consisting of answer choices and answer selection reasons (Sriyanti, Mania, & Hairani, 2019). The formative evaluation stage includes self evaluation, expert review, one to one test, and small group test. At the expert review stage, the draft instrument was validated by two validators, namely a mathematics education lecturer and a junior high school mathematics teacher. Validation aims to assess the suitability of the content, construction, and language of the question. Input from validators is used as the basis for instrument revision. Furthermore, the one-to-one stage was carried out with four students to evaluate the readability and clarity of the

questions. The last stage, namely a small group, involved 60 students to test the empirical quality of the question items.

The main instrument used in this study is the validation sheet. Validation sheets are tools used to assess the validity of diagnostic tests developed by researchers. The validation sheet in this study contains several aspects that were assessed, namely: material, construction, and language (Mulyatiningsih, 2015). The material aspect is related to the substance of algebraic materials for junior high school grade VII. The construction aspect has to do with the rules for writing good and correct test items. The language aspect is related to the use of language or sentences in each test item, guided by the General Guidelines for Indonesian Spelling. The validator team is also given space to provide suggestions, comments, and improvement recommendations on each validated test item. This will ensure validity in terms of material, construction, and language.

Data collection in this study was carried out through three main techniques: expert validation sheets, diagnostic tests, and unstructured interviews. Expert validation sheets are used to obtain an objective assessment of the feasibility of the instrument, while diagnostic tests serve to measure students' abilities and identify learning difficulties in the material being tested. Unstructured interviews are used as a supporting tool to gain more in depth information about the reasons behind students' answers and to provide context for the quantitative data obtained.

The data analysis techniques for the diagnostic tests developed include validity analysis, reliability analysis, and item analysis. The analysis of test items includes calculating the degree of difficulty, discriminating power, and effectiveness of the distractor, thus obtaining a comprehensive picture of the quality of each item in the developed diagnostic test. The results of these analyses are used as a basis for revising and refining the test items to ensure they accurately measure students' learning difficulties and misconceptions in algebra. Furthermore, this comprehensive analysis supports the development of a diagnostic instrument that is both psychometrically sound and practically applicable in classroom settings.

The assessment of the validation of the content of the instrument was carried out by experts (validators) using the rating scale presented in table 1 (Basri, Baidowi, & Turmuzi, 2021).

Table 1. Validation Criteria

Scale	Quality	Information
1	Not good	Not eligible
2	Not good	Not eligible
3	Fair enough	Underqualified
4	Good	Eligible
5	Very good	Highly qualified

Validators are required to rate each item on the instrument based on the above criteria. Each item is judged based on three aspects: material, language, and construction. Thus, there are 15 ratings for each item. The score obtained is then used to calculate the validity value of the instrument. The formula for analyzing validation data is as follows.

$$\text{Final validator value} = \frac{\text{scores obtained x 100\%}}{\text{Maximum Score}} \quad (1)$$

The interpretation of the reliability coefficient in this study refers to the criterion that a question item is declared to have high reliability if it is in the range of 0.60 to 1.00, as shown in table 2 (Basri, Baidowi, & Turmuzi, 2021).

Table 2. Reliability Coefficient Classification

(r)	Interpretation
$0.00 \leq r < 0.20$	Very low
$0.20 \leq r < 0.40$	Low
$0.40 \leq r < 0.60$	Medium/Fair
$0.60 \leq r < 0.80$	Height
$0.80 \leq r < 1.00$	Very high

According to Arikunto in Basri, Baidowi, & Turmuzi (2021), the formula for calculating reliability is as follows.

$$r11 = \left(\frac{n}{n-1}\right)\left(\frac{S^2 - \sum pq}{S^2}\right) \quad (2)$$

The interpretation of the difficulty level in this study refers to the criterion that an item is categorized as acceptable if the difficulty index is in the range of 0.3 to 0.7, as presented in table 3 (Bano, Marambaawang, & Njoeroemana, 2022).

Table 3. Difficulty Interpretation, Category

Index Distance	Interpretation
0.1 – 0.29	Difficult/difficult, bad item, revised.
0.30 – 0.70	Medium, good stuff, used
0.71 – 0.90	Easy, good, revised stuff

The formulas used to calculate the difficulty level include.

$$\text{Difficulty level} = \frac{\text{number of students who answered the questions correctly}}{\text{number of students who took the test}} \quad (3)$$

The interpretation of the force of discrimination in this study is based on the criterion that the test item is categorized as acceptable if it has a difficulty index in the range of 0.4 to 1.0, as presented in table 4 (Fitriati, 2016).

Table 4. Differential Power Interpretation Categories

Index Distance	Interpretation
0.40 – 1.00	The item was well received.
0.30 – 0.39	Items can be accepted without revision.
0.20 – 0.29	Items still need to be upgraded.
-1.00 – 0.19	Items cannot be used or need to be disposed of

This discriminatory force can be used to determine the index of differences between students with high and low ability. The discriminating strength of a test item can be determined using the following formula.

$$D = \frac{2\,(BA - BB)}{N} \quad (4)$$

Distractor refers to alternative answer choices that are not the correct answer key. Question items are said to be good if all the distractors provided can be selected proportionally by the test takers. A distractor is considered functional if it is chosen by at least 5% of all test takers.

RESULTS AND DISCUSSION

The development of diagnostic tests on algebraic materials for seventh grade junior high students was carried out using the development model proposed by Tessmer, which emphasizes formative evaluation through stages such as self evaluation, expert review, one to one, small group, and field test to

ensure the quality and effectiveness of the developed instrument. The findings of this study are as follows.

1. Introduction

The initial stage of research began with information collection through journal reviews, books, analysis of the Independent Curriculum, and review of junior high school textbooks for grade VII. In addition, an analysis of school conditions was carried out through interviews with mathematics teachers at MTsN 1 Banda Aceh. The sampling technique in this study uses purposive sampling. This technique was chosen because the research focuses on the development and initial testing of algebraic diagnostic test instruments that require subjects with specific characteristics. The research sample consisted of students in grades VII-7 and VII-10 MTsN 1 Banda Aceh who had obtained algebraic material in accordance with the learning outcomes of phase D. The selection of the class was based on considerations of curriculum suitability, affordability of the research subject, and the readiness of students to participate in all stages of instrument testing. The use of purposive sampling was assessed according to the development research objectives, which placed more emphasis on the quality of the instrument and the initial feasibility of its use rather than generalization of the research results to a wider population. The material studied is Algebra in phase D, with the learning result that students are able to express a situation in algebraic form, and use the properties of operations (commutative, associative, and distributive) to produce equivalent algebraic forms. Next, a grid of questions and a two-level multiple choice test was compiled, resulting in 35 questions summarized in the first draft. The questions developed refer to the cognitive level C1-C3. Developed only at the C1-C3 cognitive level, diagnostic tests should capture a student's cognitive foundation what they already remember, understand, and apply that is the foundation for more complex thinking skills. This restriction is carried out by taking into account the learning objectives of algebraic material in grade VII, which emphasize understanding basic concepts, the application of simple procedures, and the ability to interpret algebraic forms. In addition, the characteristics of grade VII students who are still in the early stages of learning algebra are important considerations in determining the measured cognitive depth. A focus on the cognitive level up to C3 is considered adequate to identify the conceptual and procedural difficulties students experience in the early stages of algebraic learning. Thus, this

restriction is a deliberate methodological decision to ensure the suitability between learning objectives, learner characteristics, and diagnostic test functions, rather than as a limitation of research instruments. Before assessing high level skills such as analyzing, evaluating, or creating, it is important to ensure students have adequate basic skills. If the student is not yet strong in C1-C3, then difficulties with complex skills (C4-C6) may not be due to a lack of mastery of high level thinking, but due to a poorly established foundation in memorization, comprehension, and application.

2. Self-Evaluation

The Formative Evaluation stage begins with a self evaluation, in which the researcher conducts an independent review of the prepared question design. This evaluation focuses on the suitability of the questions with the indicators and language clarity. This step aims to ensure the questions are easy to understand and ready for the expert validation stage, so that validators can focus more on assessing the content and quality of the questions. The questions developed include basic algebra material, calculating variable values in algebraic forms, simplifying algebraic forms, adding and subtracting algebraic forms, multiplying algebraic forms, dividing algebraic forms, GCF and KPK algebraic forms, and algebraic fractions.

3. Expert Reviews

The expert review stage is carried out by validating test items developed by the researchers. Validation aims to assess the quality of the instrument based on the suitability of the material, the competence measured, the indicators, the cognitive level, and the clarity of each test item. The validation process involves two validators. The first validator is a lecturer of the Mathematics Education Study Program of UIN Ar-Raniry who teaches Evaluation courses and has experience designing mathematics test items. The second validator is a mathematics teacher who is also actively helping students as a validator. He completed his undergraduate and graduate studies in the Mathematics Study Program, Learned a Bachelor of Science and a Master of Mathematics.

The validation process is carried out in three stages. In the first stage, out of the initial 35 questions, the results were as follows: 7 questions were considered appropriate, 14 questions were revised, 14 questions were deleted, and 17 new questions were added. A total of 14 questions were omitted due to

the inconsistency of the algebraic material and the existence of several similar questions. Furthermore, some new questions were added to add to the variety of questions. The analysis revealed several questions that scored above 90%. However, the item is still categorized as requiring revision due to a low score (1 or 2) from the validator on each of the 15 scoring items. Therefore, despite the high percentage score, validator feedback on improvements is still used as a reference in refining the instrument. The following is a summary of the first stage of item validation in table 5.

Table 5. Summary of Validation in the First Stage

Item	Percentage (%)	Categories	Follow-up
1	-	Invalid	Removed
2	100%	Applicable	Used
3	100%	Applicable	Used
4	97%	Invalid	Revised (there is a score of 2)
5	100%	Applicable	Used
6	100%	Applicable	Used
7	100%	Applicable	Used
8	97%	Invalid	Revised (there is a score of 2)
9	97%	Invalid	Revised (there is a score of 2)
10	97%	Invalid	Revised (there is a score of 2)
11	-	Invalid	Removed
12	-	Invalid	Removed
13	-	Invalid	Removed
14	-	Invalid	Removed
15	-	Invalid	Removed
16	93%	Invalid	Revised (there is a score of 2)
17	93%	Invalid	Revised (there is a score of 2)
18	93%	Invalid	Revised (there is a score of 2)
19	-	Invalid	Removed
20	-	Invalid	Removed
21	-	Invalid	Removed
22	97%	Invalid	Revised (there is a score of 2)
23	97%	Invalid	Revised (there is a score of 2)
24	97%	Invalid	Revised (there is a score of 1)
25	97%	Invalid	Revised (there is a score of 1)
26	-	Invalid	Removed
27	-	Invalid	Removed
28	-	Invalid	Removed
29	-	Invalid	Removed
30	100%	Applicable	Used

Item	Percentage (%)	Categories	Follow-up
31	100%	Applicable	Used
32	-	Invalid	Removed
33	97%	Invalid	Revised (there is a score of 1)
34	97%	Invalid	Revised (there is a score of 2)
35	97%	Invalid	Revised (there is a score of 1)

In the second stage, 31 questions were obtained, and 7 were disputed. The revision was made due to a small difference in the question indicator. The following is a summary of the second stage of item validation in table 6.

Table 6. Recapitulation of Validation in the Second Stage

Item	Percentage (%)	Categories	Follow-up
1	100%	Applicable	Used
2	100%	Applicable	Used
3	100%	Applicable	Used
4	100%	Applicable	Used
5	100%	Applicable	Used
6	100%	Applicable	Used
7	93%	Invalid	Revised (there is a score of 2)
8	100%	Applicable	Used
9	100%	Applicable	Used
10	100%	Applicable	Used
11	100%	Applicable	Used
12	100%	Applicable	Used
13	100%	Applicable	Used
14	100%	Applicable	Used
15	100%	Applicable	Used
16	100%	Applicable	Used
17	100%	Applicable	Used
18	100%	Applicable	Used
19	100%	Applicable	Used
20	100%	Applicable	Used
21	100%	Applicable	Used
22	100%	Applicable	Used
23	93%	Invalid	Revised (there is a score of 2)
24	93%	Invalid	Revised (there is a score of 2)
25	93%	Invalid	Revised (there is a score of 2)
26	100%	Applicable	Used
27	100%	Applicable	Used
28	100%	Applicable	Used

Item	Percentage (%)	Categories	Follow-up
29	93%	Invalid	Revised (there is a score of 2)
30	93%	Invalid	Revised (there is a score of 2)
31	100%	Applicable	Used
32	100%	Applicable	Used
33	100%	Applicable	Used
34	93%	Invalid	Revised (there is a score of 2)
35	100%	Applicable	Used
36	100%	Applicable	Used
37	100%	Applicable	Used
38	100%	Applicable	Used

In the third stage, all 38 questions were declared suitable for use as diagnostic test instruments. The questions were arranged with reference to the cognitive level C1-C3, with a C1 distribution of 8 items (21.06%), C2 of 15 items (39.47%), and C3 of 15 items (39.47%). The material and number of questions for each subject are as follows: Basic algebra (3 questions), Calculating variable values to algebraic forms (3 questions), Simplifying algebraic forms (3 questions), Adding and subtracting algebraic forms (6 questions), Multiplying algebraic forms (4 questions), Dividing algebraic forms (5 questions), GCF and KPK algebraic forms (4 questions), and Fractions of algebraic forms (10 questions).

4. One to one

This stage involves a readability test on four students. The results of the study showed that students were able to understand the intent and language of each item, so no revision was needed at this stage. Therefore, the final results show that 38 test items are considered suitable for use.

5. Small Groups

The small group phase is the final stage of this study. At this stage, the final draft resulting from the one-to-one phase was piloted on 60 students of grades VII-7 and VII-10 at MTsN 1 Banda Aceh. The test instruments were given directly to the students and then analyzed based on the level of difficulty (TK), differentiating power (DB), and the effectiveness of the distractor. The results of the trial showed that out of a total of 38 items, 22 were accepted, 10 were revised, and 6 were discarded. Based on the difficulty level analysis, 7 items were categorized as easy, 31 were categorized as

moderate, and no items were categorized as difficult. The overall reliability of the test item is 0.836, which is considered very high. Items categorized as acceptable are those that meet the criteria of difficulty, differentiating strength, and effectiveness of the distractor.

Based on the previous explanation of the research method, the difficulty level of the test item is considered acceptable if it is between 0.30 and 0.70. The discriminatory strength of a test item is considered acceptable if it falls between 0.30 and 1.00. Meanwhile, the distractor in the test item must be selected by at least 5% of the participants to be considered functional. This indicates that the item meets the good quality criteria and is suitable for use in test instruments. Acceptable test items based on the results of the analysis are presented in table 7 below.

Table 7. Accepted Test Items Based on Analysis Results

Item	Difficulty level	Differential Power	Answer Selection Functionality
2	0.700	0.387	A, B, C, D 5%≥
3	0.700	0.365	A, B, C, D 5%≥
4	0.633	0.668	A, B, C, D 5%≥
5	0.583	0.328	A, B, C, D 5%≥
6	0.517	0.347	A, B, C, D 5%≥
9	0.533	0.436	A, B, C, D 5%≥
14	0.683	0.502	A, B, C, D 5%≥
16	0.600	0.426	A, B, C, D 5%≥
17	0.683	0.486	A, B, C, D 5%≥
18	0.417	0.431	A, B, C, D 5%≥
21	0.367	0.307	A, B, C, D 5%≥
22	0.483	0.538	A, B, C, D 5%≥
24	0.433	0.381	A, B, C, D 5%≥
27	0.500	0.319	A, B, C, D 5%≥
28	0.433	0.432	A, B, C, D 5%≥
30	0.667	0.540	A, B, C, D 5%≥
31	0.450	0.524	A, B, C, D 5%≥
32	0.567	0.537	A, B, C, D 5%≥
34	0.600	0.451	A, B, C, D 5%≥
35	0.533	0.326	A, B, C, D 5%≥
36	0.467	0.606	A, B, C, D 5%≥
38	0.483	0.326	A, B, C, D 5%≥

Test items categorized as revisions are those that do not meet certain criteria for difficulty, discriminative strength, and distractor effectiveness. Test

items are considered revised if their difficulty levels are in the range of 0.71–0.90 and 0.1–0.29. Discriminatory force is considered revised if it is in the range of 0.20–0.29. Additionally, some of these items contain switchers that don't work optimally, indicated by a selection percentage of less than 5% for one or more answer choices. This condition indicates weaknesses in items that need repair. The revised test items based on the results of the analysis are presented in table 8 below.

Table 8. Revised Test Items Based on The Results of The Analysis

Item	Difficulty level	Differential Power	Answer Selection Functionality
8	0.800	0.352	A, B, C, D 5%≥
10	0.667	0.247	A, B, C, D 5%≥
15	0.817	0.497	A, B, C, D 5%≥
19	0.700	0.452	A, B, D 5%; C≥ ≤ 5%
20	0.450	0.453	A, B, C 5%; D≥ ≤ 5%
23	0.400	0.251	A, B, C, D 5%≥
25	0.867	0.224	A, B 5%; C, D≥≤ 5%
26	0.400	0.272	A, B, C, D 5%≥
29	0.833	0.439	A, B, C, D 5%≥
37	0.500	0.284	A, B, C, D 5%≥

Test items are categorized as discarded if they exhibit characteristics that are outside of the acceptance or revision criteria. Test items are discarded if the difficulty level is below 0.10 or above 0.90. Meanwhile, the discriminatory power of the test item is discarded if it is in the range of -1.00–0.19. Furthermore, the alternative answer is declared ineffective if the response to that option is less than 0.05. Test items that cannot be used (discarded) based on the results of the analysis are presented in table 9 below.

Table 9. Test Items Discarded Based on Analysis Results

Item	Difficulty level	Differential Power	Answer Selection Functionality
1	0.850	0.076	A, B, C, D 5%≥
7	0.917	0.328	C, D 5%; A, B≥≤ 5%
11	0.417	0.192	A, B, C, D 5%≥
12	0.367	0.161	A, B, C, D 5%≥
13	0.967	0.264	B 5%; A, C, D≥ ≤ 5%
33	0.433	0.138	A, B, C, D 5%≥

Based on tables 3, 4, and 5, 32 items were accepted and revised, and 6 items were rejected/discarded. A good diagnostic test can be identified

through reliability and validity testing. Before using the test, validity and reliability should be tested.

The diagnostic test items developed have undergone expert validation by validators. Qualitative validity tests are focused on the suitability indicator formulation with learning outcomes, formulated test items, test construction, options for each test item, language, and distraction analysis as described in the development process. Meanwhile, the quantitative validity test measures the level of difficulty, discrimination, and functionality of alternative answers (options) for each test item. The overall reliability test result was 0.836, which was categorized as very high.

These findings are further discussed by comparing them with relevant previous studies to examine their consistency and contribution to diagnostic assessment in mathematics learning."

The results of this study indicate that the two-tier multiple-choice diagnostic test instrument developed has characteristics that can identify students' learning difficulties and misconceptions in seventh-grade algebra material. These findings align with the results of several previous studies, which also confirm the effectiveness of the two-tier multiple-choice model as a diagnostic tool for identifying students' misconceptions. The research by Syaifuddin, Darmayanti, and Rizki (2025) developed a two tier multiple choice instrument for geometry material that proved to be valid and reliable in identifying misconceptions among junior high school students, especially regarding the concepts of triangles and quadrilaterals. This study demonstrates that, through the analysis of students' answers, certain misconceptions can be systematically identified, as was also done in this study on algebra material.

In addition, the results of research by Tukiyo, Efendi, Solissa, Yuniwati, and Pranajaya (2025) with the development of a two tier diagnostic test also found that it was effective in identifying students' misconceptions in general after undergoing validation and reliability analysis stages, with the majority of students showing misconceptions on the concepts measured. This supports the findings of this study that the two tier diagnostic test instrument is capable of revealing misconceptions, not only incorrect answers, but also the patterns of reasoning provided by students. Another study by Noprianti and Utami (2017) also found that the use of a two tier multiple choice diagnostic test, equipped with CRI, was effective in describing the percentage of students who understood the concepts, did not understand, or had misconceptions about

various science materials. Furthermore, Ningroom (2025) in his research on change of matter reinforces that two-tier diagnostic tests are capable of identifying students' misconceptions with good instrument validity. Although the context is different, this method is still relevant to the diagnostic instrument approach used in this study, demonstrating the consistency of the benefits of two tier diagnostic tests in various learning materials.

However, several studies also show the limitations of two tier diagnostic tests in distinguishing between misconceptions and lack of knowledge, so several studies have developed three tier or four tier tests to improve diagnostic sensitivity. Nevertheless, in the context of this study, the two-tier instrument developed still provides strong diagnostic indicators for identifying students' misconceptions in algebra. This study represents an initial stage in the development of a diagnostic algebra test instrument for seventh-grade students in Phase D, focusing on the measurement of cognitive levels of knowing and applying through two tier multiple choice diagnostic test accompanied by explanations. Instrument testing was limited to small groups involving relatively homogeneous subjects, due to time, resource, and funding constraints. The focus of development up to cognitive levels C1–C3 was adjusted to the learning objectives and characteristics of students in the early stages of algebra learning, so that the resulting instrument could provide an initial overview of the quality of the items and the potential of diagnostic tests in identifying students' learning difficulties. Thus, the results of this study can be used as a basis for further development and testing in a broader context and with more diverse subjects to strengthen the optimal use of instruments in mathematics learning.

CONCLUSION

This study successfully developed a two-level multiple-choice diagnostic measurement tool for seventh-grade algebra students based on Tessmer's formative research model, which demonstrated high validity and reliability. This assessment tool, designed to measure C1–C3 cognitive levels and equipped with reasoning instructions, proved effective in identifying students' learning difficulties and misconceptions, providing valuable insights for teachers to design targeted teaching and remediation strategies. Although the development was conducted on a relatively homogeneous small group and focused on Phase D algebra material, these findings provide a starting point for further refinement and wider application. Future studies are recommended

to expand this instrument to cover additional algebra topics and diverse student populations, thereby increasing its generalization and practical usefulness in mathematics learning.

DECLARATIONS

Author Contribution	:	HW: Conceptualization, Methodology, Writing-Review, Formal Analysis, and Validation MD: Writing-Review, Methodology, Formal Analysis, Validation, and Supervison
Funding Statement	:	This research was not funded.
Conflict of Interest	:	The authors declare no conflict of interest.
Additional Information	:	Additional information is available for this paper.

REFERENCES

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. http://jurnal.ideaspublishing.co.id/index.php/ideas/article/view/660#google_vignette.

Basri, K., Baidowi, J., & 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 Application*, *1*(4), 682–694. https://doi.org/10.29303/griya.v1i4.107.

Dahiana, W. O., Herman, T., Nurlaelah, E., & Pereira, J. (2023). Student semiotic representation skills in solving mathematics problems. *Jurnal Didaktik Matematika*, *10*(1), 34–47.

Duskri, M., Kumaidi, K., & Suryanto, S. (2014). Pengembangan tes diagnostik kesulitan Belajar matematika di SD. *Jurnal Penelitian dan Evaluasi Pendidikan*, *18*(1), 44–56. https://doi.org/10.21831/pep.v18i1.2123.

Fitriati, S. W. (2016). Psychological problems faced by the year eleven students of MA nuhad demak in speaking english. *English Education Journal*, *6*(1), 1–18.

McEachin, A., Domina, T., & Penner, A. (2020). Heterogeneous effects of early

algebra across California middle schools. *Journal of Policy Analysis and Management*, *39*(3), 772–800. https://doi.org/10.1002/pam.22202.

Mulyatiningsih, E. (2015). *Metode penelitian terapan bidang pendidikan*. Yogyakarta: UNY Press.

Nabilah, A., Amalia, F., Angreini, H. S., Rahmi, M., Zulkarnain, I., & Fajriah, N. (2024). Pendekatan dalam pembelajaran matematika yang dapat mengembangkan kemampuan berpikir logis dan kreatif pada siswa sekolah dasar. *Prosiding Seminar Nasional Pendidikan Matematika (SENPIKA)*, *2*, 364–372.

Ningroom, R. A. A. (2025). A two-tier multiple-choice diagnostic test to find student misconceptions about the change of matter. *Journal of Education and Learning*, 19(2), 1144-1156. https://doi.org/10.11591/edulearn.v19i2.21478.

Noprianti, E., & Utami, L. (2017). Penggunaan two-tier multiple choice diagnostic test disertai CRI untuk menganalisis miskonsepsi siswa. *Jurnal Tadris Kimiya*. https://doi.org/10.15575/jtk.v2i2.1876.

Nugraha, N., Kadarisma, G., & Setiawan, W. (2019). Analisis kesulitan belajar matematika materi bentuk aljabar pada siswa SMP kelas VII. *Journal on Education*, *1*(2), 323–334. https://doi.org/10.31004/joe.v1i2.72.

Nursalam. (2016). Diagnostik kesulitan belajar matematika: studi pada siswa SD/MI di kota makassar. *Lentera Pendidikan Jurnal Ilmu Tarbiyah dan Keguruan*, *19*(1), 1–15. https://doi.org/10.24252/lp.2016v19n1a1.

Rahayu, G., Kurniati, D., Jatmiko, D. D. H., Lestari, N. D. S., & Ambarwati, R. (2022). Analisis kemampuan berpikir kritis siswa smp dalam memecahkan masalah matematika materi bentuk aljabar ditinjau dari gaya kognitif reflektif dan impulsif. *Jurnal Edukasi dan Sains Matematika (JES-MAT)*, *8*(2), 207–216. https://doi.org/10.25134/jes-mat.v8i2.6372.

Sinabang, F. S., Lumbantoruan, N., Morina, F., Sagala, R. M., & Tanjung, S. R. (2025). Kompetensi pembelajaran aljabar atau berfikir aljabar: ekplorasi bagi calon guru matematika. *As-Salam: Journal Islamic Social Sciences and Humanities*, *3*(1), 75–85. https://ejournal.as-salam.org/index.php/assalam/article/view/90.

Singarimbun, M., & Effendi, S. (1995). *Metode penelitian survai*. Jakarta: LP3ES.

Sriyanti, A., Mania, S., & Hairani, N. (2019). Pengembangan instrumen tes

diagnostik berbentuk uraian untuk mengidentifikasi pemahaman konsep matematika wajib siswa MAN 1 makassar. *De Fermat: Jurnal Pendidikan Matematika*, *2*(1), 57–69. https://doi.org/10.36277/defermat.v2i1.40.

Sutiani, C., Herawati, L., & Hermanto, R. (2024). Pengembangan instrumen tes diagnostik untuk mengidentifikasi miskonsepsi pada materi operasi bilangan bulat. *Jurnal Karya Pendidikan Matematika*, *11*(1), 70–80. https://doi.org/10.26714/jkpm.11.1.2024.70-80.

Syaifuddin, M., Darmayanti, R., & Rizki, N. (2025). Development of a two-tier multiple-choice (TTMC) diagnostic test for geometry materials to identify misconceptions of middle school students. *Jurnal Silogisme*. https://doi.org/10.24269/silogisme.v7i2.5456.

Triyono, T., Masrukan, M., & Mulyono, M. (2023). Pengembangan tes diagnostik matematika kurikulum merdeka. *Prisma*, *12*(2), 560–569.

Tukiyo, T., Efendi, M., Solissa, E. M., Yuniwati, I., & Pranajaya, S. A. (2025). The development of a two-tier multiple-choice diagnostic test to detect students' misconceptions in learning process. *Mudir: Jurnal Manajemen Pendidikan*. https://doi.org/10.55352/mudir.v5i1.33.

Widayanti, F. D., Rahayuningsih, S., Yuliana, F., & Christian, S. (2025). Analisis kesulitan belajar matematika pada siswa kelas V MI. *Science: Jurnal Inovasi Pendidikan Matematika dan IPA*, *5*(2), 580–590. https://doi.org/10.51878/science.v5i2.5145.

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. http://jurnal.usk.ac.id/DM/article/view/17774.