Jurnal Kependidikan:
Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran
https://e-journal.undikma.ac.id/index.php/jurnalkependidikan/index

Vol. 10, No. 3 : September 2024
E-ISSN: 2442-7667
pp. 847-858
Email: jklppm@undikma.ac.id

Development of Algebra Test Using the Item Response Theory Approach for Junior High School Students

Ahmad Rustam[1]*, Ode Zulaeha[2], Wiwin Rita Sari[3], Suciati Rahayu Widyastuti[4]
[1]*Faculty of Teacher Training and Education, Universitas Sulawesi Tenggara, Indonesia.
[2]Institut Sains dan Kependidikan (ISDIK) Kie Raha Maluku Utara, Indonesia.
[3]Universitas Nahdlatul Ulama Lampung, Indonesia.
[4]Universitas Nahdlatul Ulama Cirebon, Indonesia.
*Corresponding Author. Email: ahmad.rustam1988@gmail.com

Abstract: This research aims to develop valid and reliable measuring tool for students' algebraic abilities that can be used in schools and the general public. The research follows a structured test development design, including stages such as preparing test specifications and items, field testing, revising items, and test development. The questions are aligned with the 2013 curriculum syllabus, ensuring relevance to educational standards. The test was given to 662 junior high school students in Kendari City, Indonesia, and their responses were analyzed using the item response theory (IRT) model with two logistic parameters: item difficulty level and item discriminatory power. The BILOG MG program was employed to estimate item and ability parameters. Before conducting item analysis with IRT, essential assumption tests were conducted, including unidimensional and model fit tests. The results of the development process, based on item analysis using BILOG MG, yielded 15 items covering various aspects of algebraic abilities. These items were derived from indicators such as recognizing algebraic forms, identifying elements within these forms, performing addition, subtraction, multiplication, and division operations on algebraic forms, presenting and solving real-world problems in algebraic contexts, and addressing contextual problems involving algebraic operations. The items demonstrated good fit with the model and exhibited an appropriate level of item difficulty and discriminatory power, making them suitable for use as a reliable assessment tool. Consequently, these developed tests are deemed effective for measuring students' foundational algebraic abilities.

Article History
Received: 08-06-2024
Revised: 26-07-2024
Accepted: 14-08-2024
Published: 18-09-2024

Key Words:
Test Development; Algebra Material; Item Response Theory; Two Logistic Parameters.

How to Cite: Rustam, A., Zulaeha, O., Sari, W., & Widyastuti, S. (2024). Development of Algebra Test Using the Item Response Theory Approach for Junior High School Students. *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran, 10*(3), 847-858. doi:https://doi.org/10.33394/jk.v10i3.11832

https://doi.org/10.33394/jk.v10i3.11832

Introduction

The most crucial aspect of the education sector is the evaluation procedure. Stated differently, assessment is an integral aspect of the learning activities that teachers carry out in the classroom and cannot be isolated from them. The procedure of assessment used in educational activities has a great deal of significance (Idrus, 2019). It is well-recognized that assessment exercises are crucial to the execution of curricula. One technique to determine if a learning process is effective is through evaluation (Balasubramanian et al., 2015; Mahirah, 2017; Suardipa & Primayana, 2020). An attempt is made to uphold national education quality management through evaluation activities. Evaluation is essential, and everything that has to do with education quality including evaluation must contribute to preserving that quality. In addition, assessment is a crucial task that educators complete during the teaching and learning process (Huljannah, 2021). Correct and proper evaluation can be carried out using evaluation tools that have been tested and analyzed, and this is done to assess each specific learning

achievement. During the pandemic, this evaluation process greatly facilitated the educational journey, where students were given the freedom to explore science. Additionally, of course, good and standardized test tools are needed by teachers to accurately measure students' abilities.

Currently, many test kits are available for use in evaluating student learning outcomes in classes (Elken, 2015; Sugiri & Priatmoko, 2020). An easy-to-use and often applied test tool is the multiple-choice test. However, multiple-choice tests will make a good contribution if the distractor questions work well, especially for students who demonstrate lower abilities (Budiyono, 2009). This statement aligns with the notion that the most widely used type of test tool in educational circles is the multiple-choice test (Kean & Reilly, 2014a; Raykov et al., 2019). The creation of tests is crucial because the data or outcomes they yield can accurately reflect how well students are learning in relation to their skills (Lia et al., 2020; Mohajan, 2017; Mohamad et al., 2015). Therefore, in order for educational personnel to assess pupils' true skills, proper test kits must be developed.

The classical theory used in test kits created for regular schools is highly inadequate for measuring things (Hambleton & Jones, 1993). There are currently a lot of instructors who use exam questions from published textbooks, and it is uncertain how reliable and valid these questions are. A further issue is that using traditional test theory casts doubt on the test findings due to many flaws that lead to measurement bias or variations in how the items determine students' ability (Brown, 2013; Jabrayilov et al., 2016). Embretson and Reise (2013) identified two limitations in the test: firstly, the findings of the assessment are dependent on the features of the test that is utilized; and secondly, the test takers' ability determines item factors such difficulty level and discriminating power. The fact that the tester's score is a dependent test is just one of the classical theory's many flaws. Accordingly, a test-taker may receive a higher score on a test that is simpler and a lower score on an exam that is more challenging (Subali et al., 2021). Furthermore, in classical theory, one can only search for groups not for individuals when searching for measurement error. The use of classical test theory has started to fade away with the advancement of time and science, to be replaced with contemporary theory specifically, item response theory. In order to provide more accurate measurements, this theory is predicated on two fundamental tenets: a) local independence, which refers to the possibility of answering a single item correctly with another independently, and b) unidimensionality, which denotes that the subject being measured is a single dimension (Embretson & Reise, 2013; Jabrayilov et al., 2016; Sarea & Ruslan, 2019).

The description of the issue and the significance of instrument creation in bolstering the program of autonomous learning indicate that creating customized test packages for algebraic content in junior high schools is imperative. Using a contemporary theoretical approach, namely item response theory (IRT) using two logistic parameters (2PL), namely the difficulty level parameter and the item discriminating power parameter, this development is reviewed from the perspective of item quality, such as validity and reliability (Embretson & Reise, 2013). The goal of this research is to create a viable and trustworthy instrument for assessing mathematical skills in pupils that can be utilized by both the general public and educational institutions.

Research Method

The research method employed was research and development, collecting response data from 662 junior high school students in Kendari City, Indonesia. The response data were

used to analyze the development of algebra material test instruments, which can later be utilized by junior high school students and others needing the test.

Research Stages

The stages of the research followed the test development steps outlined by Hambleton & Jones (1993) and Irvine & Kyllonen (2013):

Preparation of Test Specifications

The process begins with identifying algebraic material based on competency standards, basic competencies, and indicators from the 2013 curriculum. The analysis was conducted descriptively, presenting a grid of test instruments to be developed.

Preparation of the Test Item Pool

The researcher analyzed various relevant references to develop the test items. Based on the existing material indicators, items were developed for each indicator.

Field Testing the Items

After arranging the items into a test package, this stage aimed to determine whether the test instructions could be understood properly and whether the items did not present ambiguous instructions. This stage was conducted on a small group sample, consisting of one class of 50 students.

Revision of the Test Items

Items that received student responses were analyzed based on student response patterns using IRT analysis, reviewing the question sentences, answer keys, and item distractors. This analysis utilized the theory of item responsiveness by examining the differential power values of items (a) and the difficulty level of items (b).

Test Development

In this stage, data were collected in the field using a large sample. The test was conducted on 662 junior high school students in Kendari City. After obtaining the response data, the data were analyzed using BILOG MG software with IRT, specifically the two logistic parameters. Before item analysis using IRT, prerequisite tests such as the unidimensional test and the model fit test were conducted. The BILOG MG program in IRT analysis can identify 1 to 3 item parameters (Alkursheh et al., 2022). The results of this study included the analysis of the test items, identifying the parameters of the differential power of the items (a) and the difficulty level of the items (b). This analysis provided data about the algebraic abilities of each student.

Results and Discussion

Preparation of Test Specifications

Exam requirements preparation begins with determining algebraic content for junior high school students in Class VII by consulting indications, fundamental skills, and competence criteria from the 2013 curriculum. When determining indicators for test item creation, a number of fundamental capabilities and resources are consulted. Table 1 lists these in detail, along with the fundamental abilities and question indicators that go along with them.

Table 1: Basic Competencies and Indicators

No.	Basic competencies	Question Indicator
1	Explaining algebraic forms and their elements using contextual problems	1. Recognize algebraic forms 2. Identify the elements of algebraic forms

No.	Basic competencies	Question Indicator
2	Explain and perform operations on algebraic forms (addition, subtraction, multiplication, and division)	1. Solve the addition and subtraction operations of algebraic forms 2. Solve the multiplication operation of algebraic forms 3. Solve algebraic division operations
3	Solve problems related to algebraic forms	1. Presenting real problems in algebraic form 2. Solve algebraic forms in real problems
4	Solve problems related to operations on algebraic forms	1. Solve contextual problems on the operation of algebraic forms 2. Solve real problems on the operation of algebraic forms

Curriculum Analysis and Test Item Development

Based on the curriculum analysis for Grade VII Middle School students, several basic competencies were identified to assess algebraic abilities, including:

- Explaining algebraic forms and their elements using contextual problems.
- Explaining and performing operations on algebraic forms (addition, subtraction, multiplication, and division).
- Solving problems related to algebraic forms.
- Solving problems related to operations on algebraic forms.

Nine indications were found in the grid. Nevertheless, three items were created in various formats for indicators 1 and 2, in addition to additional indications, making a total of 15 items for the exam. The algebraic ability exam for junior high school students in Class VII was designed using the test item grids, which were created in conjunction with the subject instructors. The Middle School (SMP) Grade VII curriculum study has led to the identification of many fundamental competencies that will be used to assess the algebraic proficiency of these pupils. These competencies include the ability to perform addition, subtraction, multiplication, and division operations on algebraic forms, solve problems pertaining to algebraic forms, and solve problems pertaining to operations on algebraic forms. They also include the ability to explain algebraic forms and their elements using contextual problems. This investigation yielded nine indications. There were three distinct items created for indicators 1 and 2, in addition to additional indications, for a total of 15 items in the exam. The algebraic ability exam for pupils in Class VII was designed using the test item grids, which were created in conjunction with the subject instructors.

Preparation of the Test Item Pool

In order to create test items, this stage entails examining multiple sources. Test items were developed for each indication in the grid based on the previously identified question indicators. Before creating the questions, a number of factors were taken into account, particularly in relation to certain elements that required attention. According to Muhsetyo et al. (2014), algebraic content is frequently challenging for pupils to understand. These challenges may lead to mistakes while responding to test questions; this is corroborated by Lord & Novick's (1968) finding that a number of variables, such as the degree of item complexity, may affect exam question errors. Students' mistakes are proof of the challenges they have in understanding the subject matter.

Learning difficulties in mathematical content can be interpreted as challenges students face, which can be observed from the pattern of errors made while solving problems (Kereh

et al., 2013). According to Mardianto (2012), the factors causing learning difficulties can be broadly categorized into:

Internal factors - Issues or conditions arising from within the students.

External factors - Issues or conditions originating from outside the students.

Given that algebraic material contains abstract concepts, these concepts need to be reinforced continuously. The results of previous research provide a basis for developing test items.

Development of Algebra Material Items

In order to produce the algebra material objects, question cards have to be made. This method ensures that each item consists of a single card by helping to identify its features. Every question card aims to clarify the question's identification. Particularly, every test item card includes information on subjects, courses, semesters, curriculum kinds, resources, question indicators, cognitive levels or dimensions, content descriptions, answer keys, and standards for test outcomes.

Field Testing the Items

The goal of the field-testing phase is to ascertain whether the test instructions are understandable and whether the items don't have confusing instructions. Following the compilation of the test items into a test package, this step is carried out. This phase also includes testing the things in small groups—that is, in a class of fifty pupils. Students from SMPN 5 Kendari who had studied algebraic principles made up the responders.

Parameters of Test Items

In item response theory (IRT), three types of item parameters are identified:

1) Discrimination parameter (a) - This measures the ability of an item to differentiate between respondents with different levels of ability.
2) Difficulty parameter (b) - This indicates the level of difficulty of the item.
3) Guessing parameter (c) - This represents the likelihood of a low-ability examinee guessing the item correctly.

These parameters are used in various logistic models:

- Three-parameter logistic model (3PL)**:** Includes parameters for discrimination (a), difficulty (b), and guessing (c).
- Two-parameter logistic model (2PL)**:** Includes parameters for discrimination (a) and difficulty (b), with the guessing parameter (c) assumed to be zero (c = 0).
- One-parameter logistic model (1PL)**:** Includes only the difficulty parameter (b), with the discrimination parameter (a) assumed to be constant (a = 1) and the guessing parameter (c) assumed to be zero (c = 0).

Table 2. Difficulty of small group test items

No.	Item Difficulty Index (b_i)	Category	Amount	Item Number
1	$b_i > +2$	Hard	1	15
2	$-2 \leq b_i \leq +2$	Medium	12	1,2,3,4,5,7,9,10,11,12, 13,14
3	$b_i < -2$	Easy	2	6,8
Total Items			15	

The analysis presented in Table 2 provides important insights into the difficulty levels of the test items. It was found that 80% of all items fall within a moderate or good level of difficulty, indicating that the majority of the test items are appropriately challenging and effectively measure student abilities. Only a very small proportion, 0.07% of all items, are categorized as difficult, which still contributes positively to the overall assessment. Based on these findings, the researchers concluded that approximately 87% of all test items

successfully reflect the students' capabilities, as they span moderate to difficult levels of difficulty. However, 13% of the test items were identified as easy and were classified as poor items. These items are considered less effective in evaluating student performance, suggesting that they might not sufficiently challenge the students or differentiate between varying levels of ability.

Parameters of Grain Difference (a)

For each specific ability scale, the grain's degree of slope at the item difficulty point is described by the grain differential power (ai) parameter. Researchers elucidated that the differential power increases with the slope of the curve. The theoretical range for the grain power parameter values, according to Pyrczak (1973), spans from between $-\infty \leq a_i \leq \infty$. However, in practical applications, the differential power parameter values are typically observed within a more restricted range of $0 \leq a_i \leq 2$. In this study, the grain discrepancy parameter analysis was conducted using the BILOG MG program, which generated specific output data. Upon analyzing the values in the slope column from this output, the researchers identified that the differential power parameter for this study is 0.896. The classification of these grain power parameters, as determined from the analysis, is summarized in Table 3, providing a detailed view of how this parameter fits within the broader context of the test item characteristics.

Table 3. Differential power of items in small group trials

No.	Different Power (a_i)	Category	Amount	Item Number
1	$a_i > 2$	Not good	-	
2	$0 \leq a_i \leq 2$	Good	15	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
Total Items			15	

Table 3 shows that approximately all grains have appropriate specifications for discriminating power. Researchers came to the conclusion that all test items 100% of them were able to characterize a good function in terms of giving information about students' skills. Conversely, 0% of the test items show a low score for discriminating power.

Revision of the Test Items

In this stage, the test items that received student responses are analyzed based on the response patterns. Researchers review the wording of the test item questions, answer keys, and distractors. The analysis was conducted using Item Response Theory (IRT) to determine the differential power parameters (a) and item difficulty levels (b) (Kean & Reilly, 2014b). The results of the analysis indicated that approximately 87% (13 out of 15) of the items demonstrated a good level of difficulty. Researchers concluded that these items effectively assessed students' abilities. However, 13% (2 out of 15) of the items, specifically items 6 and 8, were categorized as easy and did not meet the desired difficulty level. Item 15 was found to contain an error in the answer key, resulting in most students answering incorrectly. Additionally, items 6 and 8 required revisions due to overly simplistic mathematical language. Despite these issues, the researchers included the problematic items in a large-scale test, assuming that a larger sample size would yield a broader range of responses.

Test Development

At this stage, data collection was conducted in the field with a large sample. Tests were administered to 662 students from SMP Negeri 5 Kendari in the Kolaka district. The response data were analyzed using IRT with two logistic parameters (2PL), utilizing BILOG MG software. This analysis provided insights into the differential power parameters (a) and item difficulty levels (b) for each test item, ultimately generating data on students' algebraic abilities (Zimowski, 2017).

The following sections present the results and discussion of the test item analysis:

- Differential Power Parameters (a): These parameters measure how well an item discriminates between students of different ability levels.
- Item Difficulty Levels (b): These parameters indicate the difficulty level of each item, ranging from very easy to very difficult.

The analysis revealed that most items were appropriately challenging and capable of distinguishing between students with varying levels of algebraic proficiency. The results underscore the importance of ongoing item analysis and revision to ensure the validity and reliability of the assessment tool.

Test the Assumptions of IRT

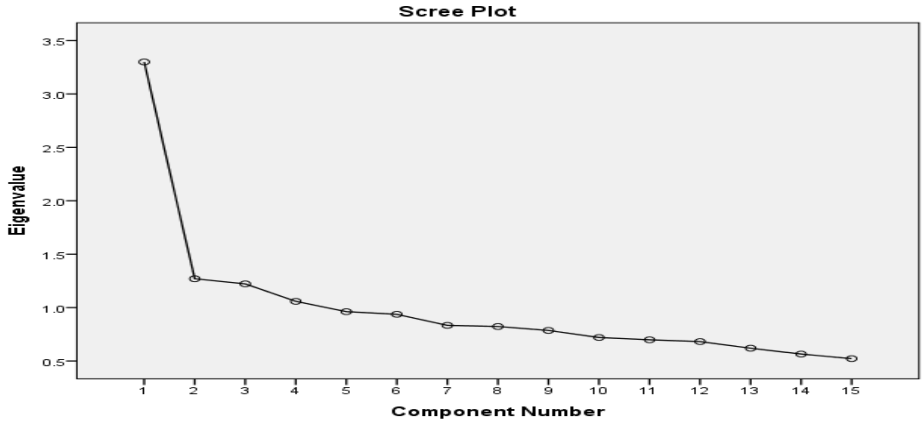
Unidimensional test

The objective of factor analysis item analysis outcomes is to ascertain the unidimensional test. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) score is 0.805 and is larger than the 0.05, and the Chi-squared value on the Barlet test is 1175.524 with 105 degrees of freedom and a p-value of 0.000 based on the study findings. The table below includes all of the following information in full:

Table 4. Unidimensional Test

KMO and Bartlett's Test		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy.		.805
Bartlett's Test of Sphericity	Approx. Chi-Square	1175.524
	df	105
	Sig.	.000

Additionally, the unidimensional item analysis includes a review of eigenfactors, which are the factors formed from the data. These factors demonstrate eigenvalues that are relatively consistent, ranging from 3 to 1. This indicates that the measurement primarily assesses one specific dimension. Therefore, researchers concluded that the test instrument is unidimensional, aligning with findings by Brown and Moore (2012) and Hox (2021) regarding dominant dimensions in measurement. This conclusion is further supported by Hambleton and Swaminathan (1985), who also affirm the presence of a dominant dimension, confirming the unidimensional nature of the test set (Hambleton & Swaminathan, 1985).



Model Fit Test

To ascertain if the items are appropriate and whether more research is required, the model fit test analysis is essential. This evaluation makes sure the objects match the specified model. The model fit significance value for each of the 15 components was determined using BILOG MG. The findings showed that all of the items matched the model since the significant value for the Threshold column (item difficulty level) of each item was greater than the threshold value, $\alpha = 0.05$.

Evaluation of Approximation Findings Based on Item and Ability Parameters Item parameters and ability parameters are two important factors that are calculated in item response theory (IRT). Theta, or "θ," the ability parameter, describes test-takers according to their ability level θ. Item parameters, on the other hand, use the logistic model to characterize item attributes. The differential power parameter, item difficulty level, and guess parameter are some of these item parameters. The three-parameter logistic model incorporates differential power (a), difficulty level (b), and guess (c) parameters. The two-parameter logistic model includes differential power (a) and difficulty level (b), with an assumed guess parameter value of zero (c = 0). Lastly, the one-parameter logistic model only includes the difficulty level parameter (b), with a constant differential power parameter value of 1 (a = 1) and a zero-guess parameter value (c = 0). These parameter estimations are crucial for accurately assessing both item characteristics and respondent abilities within the IRT framework.

Item Difficulty Level Parameter (b)

The item difficulty level parameter, denoted as bi, is designed to assess students' abilities by measuring how well they can answer items across various difficulty levels. This parameter functions as a reflection of a person's ability to tackle specific items. In theory, based on grain response theory, the item difficulty level parameter bi can range from $-\infty \leq b_i \leq \infty$. However, in practical applications, this range is usually more constrained, typically spanning from $-2 \leq b_i \leq +2$. Items with a difficulty level below −2 are classified as easy and fall into the low-difficulty category. Items with a difficulty level above +2 are classified as difficult. The analysis of the BILOG MG output for this study indicates that the item difficulty level ranges from −1.579 to 1.591. This suggests that all items analyzed fall within a practical and moderate range of difficulty, neither too easy nor too difficult, thereby providing a balanced assessment of student abilities. The classification of these item difficulty parameters, based on the output, is summarized and presented in the subsequent table.

Table 5. Difficult Level for large group test items

No.	Item Difficulty Index (b_i)	Category	Amount	Item Number
1	$b_i > +2$	Hard	0	
2	$-2 \leq b_i \leq +2$	Medium (Good)	15	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
3	$b_i < -2$	Easy	0	-
Total Items			15	

Table 5 shows that all test items have a satisfactory degree of difficulty, which means that every item is suitably demanding and accurately assesses students' skills. The researchers came to the conclusion that all test items had the ability to correctly reflect students' skills based on these findings. Furthermore, the study shows that, in terms of difficulty level, 0% of the test items are classified as poor items. This implies that none of the test's items are either simple or very complex, demonstrating the test's suitability and balance for evaluating students' skills across the specified difficulty spectrum.

Parameters of Grain Difference (a)

On an ability scale, the slope of the curve at the item difficulty point is represented by the grain differential power parameter (ai). This slope is significant because it shows how a student's skill level affects the chance of a right response, changing it dramatically. The differential power increases with slope steepness, indicating that the item is more successful in differentiating across students with varying skill levels. Conceptually, the differential

power parameter (a_i) can theoretically range between $-\infty \leq a_i \leq \infty$. However, in practical applications, the values typically observed for this parameter are within a more limited range, from $0 \leq a_i \leq 2$. This range is more commonly used in practice to assess the effectiveness of test items. In this study, the analysis of the differential power parameter was conducted using the BILOG MG program, with the results displayed in the program's output. The analysis of the slope column from the output indicates that the differential power for the items is 1.021. This value falls within the practical range and suggests that the items have a good ability to differentiate between students of varying ability levels. The detailed classification of the grain power parameters based on this analysis is presented in the subsequent table.

Table 6. Difficult Level for large group test items

No.	Different Power (a_i)	Category	Amount	Item Number
1	$a_i > 2$	Not good	0	-
2	$0 \leq a_i \leq 2$	Good	15	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
Total Items			15	

Tabel 6 menunjukkan bahwa 100% dari semua butir soal (item tes) menunjukkan daya pembeda yang baik. Ini berarti bahwa setiap item tes secara efektif membedakan antara siswa dengan tingkat kemampuan yang berbeda. Berdasarkan temuan ini, para peneliti menyimpulkan bahwa 100% dari semua item tes secara akurat mencerminkan kemampuan siswa. Lebih jauh, analisis menunjukkan bahwa 0% dari item tes memiliki daya pembeda yang buruk, yang menyiratkan bahwa tidak ada item yang tidak efektif dalam membedakan antara siswa berdasarkan kemampuan mereka. Ini menunjukkan bahwa item tes tersebut disusun dengan baik dan berkinerja baik dalam menilai kemampuan siswa di seluruh spektrum kemampuan.

Test Participants Ability Parameters (θ)

One metric that displays the qualities of the test takers' abilities is the ability parameter (θ). The output of the BILOG MG program's analysis findings displays the test taker's ability estimation results. Regarding the output outcomes, the empirical dependability value is 0.7182 and the average value of the students' talents is −0.0006. The average student ability number, which is negative, shows that most pupils typically display ability (θ), which is still rather low.

Discussion

This study contributes new understanding in the development of algebra ability assessment tools using item analysis based on BILOG MG. Some conceptual implications of the results of this study include: the Validity and Reliability of Assessment Tools, this study shows that by using BILOG MG, test items can be evaluated comprehensively to ensure their suitability with the model used. This strengthens the concept that the validity and reliability of a test are greatly influenced by the quality of its items. Comprehensive Algebra Ability Measurement, by covering various aspects of algebraic ability such as recognition of algebraic forms, basic operations, and contextual problem solving, this study supports the view that a good test should cover various aspects of the ability to be measured. Item Response Model (IRT), this study confirms the superiority of IRT, especially the two-parameter model (2PL) in BILOG MG, in evaluating test items. This supports the concept that IRT is superior to classical test theory in providing detailed information about the level of difficulty and discriminatory power of items.

The results of this study have several important practical implications for the development and use of assessment tools in schools, including the Development of More Effective Assessment Tools. By producing valid and reliable items, the developed test can be

used effectively to measure students' basic algebra skills. This means that teachers and test developers can use the test with confidence that the results accurately reflect students' abilities. Learning Enhancement and Assessment, items that cover various aspects of algebra skills allow teachers to get a comprehensive picture of students' abilities. This can be used to design more targeted and effective learning interventions. Formative and Summative Assessment, the test can be used both in formative assessment to monitor student progress and in summative assessment to evaluate final achievement. Items that have good discriminating power can help identify students who need additional help or who have higher abilities.

Through this developmental process, researchers observed that many students struggle with solving contextual problems in algebraic operations. Particularly, issues arise with understanding algebraic laws such as the distributive and associative laws, alongside challenges with variable coefficients in algebraic forms. Although students grasp concepts like the identity law of addition and multiplication, as well as terms and constants, there is room for improvement. The test's existence offers educators insights into specific areas where students may need remedial or enrichment support from an early stage. This strategic approach allows for tailored interventions that align with individual learning needs. Test development is crucial for providing accurate and valid descriptions of student learning outcomes based on their abilities (Lia et al., 2020; Mohajan, 2017; Mohamad et al., 2015). It serves as a rational tool for informed decision-making, enhancing quality control in education. The continuous development and refinement of tests are highly anticipated in the education sector to ensure effective learning assessments and support student progress.

Conclusion

The results of this study conclude that based on item analysis using BILOG MG, yielded 15 items covering various aspects of algebraic abilities. These items were derived from indicators such as recognizing algebraic forms, identifying elements within these forms, performing addition, subtraction, multiplication, and division operations on algebraic forms, presenting and solving real-world problems in algebraic contexts, and addressing contextual problems involving algebraic operations. The items demonstrated good fit with the model and exhibited an appropriate level of item difficulty and discriminatory power, making them suitable for use as a reliable assessment tool. Consequently, these developed tests are deemed effective for measuring students' foundational algebraic abilities.

Recommendation

For future research, identifying test items that indicate differential item function for gender groups is necessary. In addition, using the Rasch model approach as a comparison of results is essential. When creating Algebra operations questions, teachers must be careful in writing or using brackets in operations.

Acknowledgements

Directorate General of Higher Education Ministry of National Education of the Republic of Indonesia, which funded this research assignment. Furthermore, the University of Southeast Sulawesi through the Chancellor and research institute, community service as a facilitator providing this research assignment.

References

Alkursheh, T. O., Al-zboon, H. S., & AlNasraween, M. S. (2022). The Effect of Item Form on Estimating Person's Ability, Item Parameters, and Information Function According to Item Response Theory (IRT). *International Journal of Instruction*, *15*(3), 1111–1130.

Balasubramanian, B. A, Cohen, D. J, Davis, M. M, Gunn, R, Dickinson, L. M, Miller, W. L, ... & Stange, K. C. (2015). Learning evaluation: blending quality improvement and implementation research methods to study healthcare innovations. *Implementation Science*, *10*(1), 1–11.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In *Handbook of structural equation modeling* (p. 361,379).

Brown, J. D. (2013). *Classical test theory* (pp. 337–349). n The Routledge handbook of language testing.

Budiyono. (2009). The Accuracy of Mantel-Haenszel, Sibstest, and Logistic regression Methods in Differential Item Functioning Detection. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(1), 1–20.

Davier, M. Von, Yamamoto, K., Shin, H. J., Chen, H., & Khorramdel, L. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 466–488.

Elken, M. (2015). Developing policy instruments for education in the EU: The European qualifications framework for lifelong learning. *International Journal of Lifelong Education*, *34*(6), 710–726.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Hakim, M. L., Muslim, & Ramalis, T. R. (2019). Karakteristik Tes Hasil Belajar Ranah Kognitif Materi Elastisitas Menggunakan Analisis Item Response Theory. *Jurnal Penelitian Pembelajaran Fisika*, *10*(1), 22–32.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement*, 38–47.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory Principles and Applications*. Springer Science+Business Media, LLC.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE Publications Inc.

Hox, J. J. (2021). Confirmatory factor analysis. *The Encyclopedia of Research Methods in Criminology and Criminal Justice*, *2*, 830–832.

Huljannah, M. (2021). Pentingnya Proses Evaluasi dalam Pembelajaran di Sekolah Dasar. *Directory of Elementary Education Journal*, *2*(2), 49–63.

Idrus, L. (2019). Evaluasi Dalam Proses Pembelajaran 1. *Evaluasi Dalam Proses Pembelajaran*, *9*(2), 920–935.

Irvine, S. H., & Kyllonen, P. C. (2013). *Item generation for test development* . Routledge.

Jabrayilov, R, Emons, W. H, & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572.

Kean, J., & Reilly, J. (2014a). Item response theory. In *Handbook for clinical research: Design, statistics and implementation* (pp. 195–198).

Kean, J., & Reilly, J. (2014b). Item response theory. In *Handbook for clinical research: Design, statistics and implementation* (pp. 195–198).

Kereh, C. T., Sabadar, J., & Tjiang, P. C. (2013). Identifikasi kesulitan belajar mahasiswa dalam konten matematika pada materi pendahuluan fisika inti. *Proceedings of*

Seminar Nasional Sains Dan Pendidikan Sains VIII, Fakultas Sains Dan Matematika, UKSW Salatiga, *4*(1), 11–12.

Lia, R. M, Rusilowati, A, & Isnaeni, W. (2020). NGSS-Oriented chemistry test instruments: validity and reliability analysis with the rasch model. *REiD (Research and Evaluation in Education)*, *6*(1), 41–50.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mahirah, B. (2017). Evaluasi belajar peserta didik (siswa). *Idaarah: Jurnal Manajemen Pendidikan*, *1*(2).

Mardianto. (2012). *Psikologi Pendidikan*. Perdana Publishing.

Mohajan, H. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University Economics Series*, *17*(4), 59–82.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia-Social and Behavioral Sciences*, *204*, 164–171.

Muhsetyo, G., Krisnadi, E., & Wahyuningrum, E. (2014). *Pembelajaran matematika SD*. Universitas Terbuka.

Pyrczak, F. (1973). Validity of the Discrimination Index As A Measure of Item Quality. *Journal of Educational Measurement*, *10*(3), 227–231.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & Harrison, M. (2019). On true score evaluation using item response theory modeling. *Educational and Psychological Measuremen*, *79*(4), 796–807.

Retnawati, H. (2013). Pendeteksian Keberfungsian Butir Pembeda dengan Indeks Volume Sederhana berdasarkan Teori Respons Butir Multidimensi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 275–286.

Sarea, M. S., & Ruslan, R. (2019). "Karakteristik Butir Soal: Classical Test Theory vs Item Response Theory?." *Didaktika: Jurnal Kependidikan 13.1*, *13*(1), 1–16.

Sirodj, D. A. N. (2018). Analisis Kualitas Aitem Intelligence Structure Test (IST) melalui Metode Item Response Theory (IRT). *Schema: Journal of Psychological Research, 4*(2), 98–107.

Soedjadi, R. (1996). Diagnosis Kesulitan Siswa Sekolah Dasar dalam Belajar Matematika. *Jurnal Jurusan Matematika FPMIPA IKIP Surabaya*, 25–33.

Suardipa, I. P., & Primayana, K. H. (2020). Peran desain evaluasi pembelajaran untuk meningkatkan kualitas pembelajaran. *Widyacarya: Jurnal Pendidikan, Agama Dan Budaya*, *4*(2), 88–100. http://mpoc.org.my/malaysian-palm-oil-industry/

Subali, B., Kumaidi, & Nonoh, S. A. (2021). The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory. *International Journal of Instruction*, *14*(1), 647–660.

Subali, B., Kumaidi, Nonoh, S. A., & Sumintono, B. (2019). Student achievement based on the use of scientific method in the natural science subject in elementary school. *Jurnal Pendidikan IPA Indonesia*, *8*(1), 39–51.

Sugiri, W. A., & Priatmoko, S. (2020). Persprektif asesmen autentik sebagai alat evaluasi dalam merdeka belajar. *At-Thullab: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *4*(1), 53–61.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171.

Zimowski, M. F. (2017). BILOG-MG. In *Handbook of Item Response Theory* (pp. 435–446). Chapman and Hall/CRC.