Jurnal Math Educator Nusantara
Wahana publikasi karya tulis ilmiah di bidang pendidikan matematika
p-issn: 2459-9735 e-issn: 2580-9210
http://ojs.unpkediri.ac.id/index.php/matematika

Analysis of the characteristic of mathematical literacy test items on algebraic elements using IRT

Jasmine Nurul Izzah [1] , Sintha Sih Dewanti[2] *

[1,2]Program Studi Pendidikan Matematika, Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
Jl. Laksda Adisucipto, Papringan, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia.
E-mail: sintha.dewanti@uin-suka.ac.id *

Article received : August 29, 2025.
Article revised : January 14, 2026.
Article Accepted : January 24, 2026.
Article Publish : May 20, 2026
** Corresponding author*

Abstract: Accurate measurement tools are needed to assess mathematical literacy in algebra. This study purposes to describe the test items characteristics in a mathematical literacy test using IRT. This study conducted by modifying the test development steps according to Downing & Haladyna. Subjects in this study were 179 7th students at a junior high school in Yogyakarta. The data collection instruments included a validation sheet and a test instrument. Data analysis techniques included content validity, construct validity, construct reliability, item and person characteristics using the PCM model. The results of the study indicate that there are eight items of mathematical literacy in algebra for seventh grade. All items are content-valid with a value of V >0.80, construct-valid (SLF > 0.45), and construct-reliable (CR>0.70). The difficulty level of items is in the good category with a b value in the range -2<b<2. The ICC curve shows that there are 2 items with ideal ICC and 6 items with non-ideal ICC. The information function and SEM indicate that the test instrument provides the best information when administered to students with abilities slightly below average. There were 6 questions that met the item fit criteria and 112 students who met the person fit criteria.

Keywords: algebra; mathematical literacy; item response theory

INTRODUCTION

Mathematical literacy refers to an individual's ability to reason mathematically and formulate, use, and interpret mathematics to solve problems in various real-world contexts (O.e.c.d, 2023b). Simply put, mathematical literacy is the ability to understand and apply mathematics in everyday life (Ojose, 2011). Mathematical literacy helps students connect mathematics with real-world problems in diverse situations. With mathematical literacy, they can think mathematically and solve problems using the mathematical concepts they have learned. This means that mathematical literacy plays a very important role for students.

At the international level, mathematical literacy skills are measured through the Programme for International Student Assessment (PISA) conducted by the OECD. Mathematical literacy skills in PISA are measured based on three components, including process, content, and context. The process component refers to the steps a person takes to solve problems using mathematics (Zahrah, 2024), including formulate, employ, interpret, and reasoning. Mathematical content is the component defined as the subject matter or topics of mathematics studied in school (Zahrah, 2024), consisting of quantity, uncertainty and data, change and relationships, and space and shape. The context component describes the

CITATION FORMATS: Izzah, J. N. ., & Dewanti, S. S. (2026). Analysis of The Characteristic of Mathematical Literacy Test Items on Algebraic Elements Using IRT. *Jurnal Math Educator Nusantara: Wahana Publikasi Karya Tulis Ilmiah Di Bidang Pendidikan Matematika*. 12(1), 16-31. https://doi.org/10.29407/jmen.v12i1.26949

problems being tested, including personal context, work context, social context, and scientific context.

The mathematical literacy scores of Indonesian students in PISA 2022 are still unsatisfactory. Compared to the OECD average of 472, the average score of Indonesian students was 366, with the lowest average score in the area of change and relationships. In the Indonesian curriculum, the content area of change and relationships is closely related to the subject matter in the curriculum, namely algebra (Farida et al., 2021). The low level of mathematical literacy in this field is also reflected in research conducted by (Selan et al., 2020). Many students are only able to solve problems up to the stage of creating a model and applying the model design, but they still struggle to find the correct solution and interpret it in a real-world context.

Algebra is one of the subjects in the school curriculum that is widely applied to solve real-life problems. Algebra is first introduced to seventh-grade students after they have studied arithmetic in elementary school. When learning algebra, students begin to experience a significant change in their thinking process, shifting from arithmetic thinking to algebraic (abstract) thinking (Ardiansari, 2018). Therefore, at this stage, students need to fully understand the transition from using only numbers to numbers and letters. The algebraic elements taught in seventh grade include algebraic expressions, as well as linear equations and inequalities with one variable.

Regarding mathematical literacy skills, Wibowo et al. (2020) mention that the questions used by teachers in learning generally only test routine procedures, are unable to relate mathematical contexts to everyday problems, and are not in line with mathematical literacy indicators. To determine students' mathematical literacy skills, particularly in algebra, teachers need a testing instrument. A test instrument can be considered a good measuring tool if it fulfills the validity and reliability requirements (Mardapi, 2017). The testing instrument used must be developed in accordance with testing procedures and accompanied by item analysis to ensure accurate data.

Item analysis is conducted to determine whether the items that make up a test instrument are capable of functioning as adequate measuring tools or not (Siregar et al., 2024). Item analysis can be performed through Classical Test Theory (CTT) or Item Response Theory (IRT). CTT is a widely used test theory. The same applies to estimates of students' mathematical literacy abilities, which have been based on the results of global response analysis using CTT (Dewanti et al., 2021). However, Hambleton et al. (1991) stated that CTT has a fundamental weakness, which is the inability to separate test taker characteristics from test characteristic. This means that item characteristics can change according to test taker characteristics and vice versa, making it difficult to generalize to other test takers.

To solve the weakness of CTT, IRT was developed, which is more advantageous in estimating both item parameters and test parameters (Santoso, 2018). IRT was developed to overcome CTT's lack of independence from the groups of participants taking the test and from the test items being tested (Sarea & Ruslan, 2019). IRT develops a model that links item characteristics to individual characteristics (Dewanti et al., 2021). Therefore, the results of

analysis using IRT will be more accurate than the results of CTT, which still depend on the sample.

IRT was developed based on several principles, including: 1) an individual's test results can be expected based on their abilities, and 2) The relationship between test scores and abilities is described by a function known as the item characteristic curve (ICC) (Hambleton et al., 1991). The ICC describes the relationship between a participant's ability level (θ) and the proportion of correct answers P_i (θ) (Sumaryanta, 2021). The probability of answering correctly will be greater for individuals with higher levels of ability.

Hambleton et al. (1991) state that there are three assumptions in item response theory, namely unidimensionality, local independence, and parameter invariance. Unidimensionality means that only one ability is measured by the test item (Retnawati, 2014). The unidimensionality assumption is satisfied if the measurement results show that the dominant dimension is only one (Sudaryono, 2012). The second IRT assumption is local independence, which is a condition where, if the factors influencing performance are constant, then the subject's responses to any pair of items will be statistically independent of each other (Retnawati, 2014). When the unidimensionality assumption is met, the local independence assumption is also satisfied (Hambleton et al., 1991). The third assumption in IRT is parameter invariance of items and ability parameters. The assumption of parameter invariance indicates that item parameters are independent of the ability distribution of test takers, while test-taker parameters are independent of the set of items used (Hambleton et al., 1991).

IRT can be used to analyze both dichotomous and polytomous item scoring. Dichotomous scoring is typically used for multiple-choice questions, where correct answers are given a score of 1 and incorrect answers are given a score of 0, while polytomous scoring is used for items with multi-category responses, such as Likert scales for attitude scales and essay test responses. There are three popular unidimensional IRT models: the one-parameter logistic model, the two-parameter logistic model, and the three-parameter logistic model, abbreviated as 1PL, 2PL, and 3PL, respectively, which are suitable for item response data with dichotomous scoring (Hambleton et al., 1991). The 1PL model uses a single parameter, namely item difficulty, the 2PL model uses two parameters, namely item difficulty and discrimination, while the 3PL model uses three parameters, namely difficulty, discrimination, and pseudo-guessing (Hambleton et al., 1991). The polytomous IRT models commonly used are the Graded Response Model (GRM), Partial Credit Model (PCM), and Generalized Partial Credit Model (GPCM) (Retnawati, 2014).

Several previous studies have developed mathematical literacy instruments on various mathematical materials and elements, but generally have not focused on algebraic elements in seventh grade. For example, research conducted by Wibowo et al. (2020), developed mathematical literacy instruments on cubes and blocks, while Apriatni et al. (2022) developed mathematical literacy instruments on high school trigonometry. In addition, research by Cendana et al. (2024) developed mathematical literacy instruments using the context of Krakatoa Cilegon Batik, and other research. Based on previous studies, there have not been

many studies that specifically develop and analyze mathematical literacy instruments on algebraic elements, especially in seventh grade. Therefore, this study aims to fill this gap.

Most of the time, item analysis on mathematical literacy test instruments is still conducted using CTT, which is less capable of accurately explaining students' abilities. Some examples are studies conducted by (Cendana et al., 2024; Nursakiah et al., 2022; Wibowo et al., 2020). Therefore this study purposes to describe the characteristics of test items in the seventh-grade algebraic mathematical literacy test instrument developed by the researcher using IRT. The use of IRT for item analysis is expected to produce a higher quality mathematical literacy test instrument for seventh-grade algebra elements. Additionally, this study is expected to enrich the study of mathematical literacy test instrument development and the application of IRT in item analysis.

METHODS

This study uses the Research & Development (R&D) method with test development procedures by Downing & Haladyna (2006) modified into 10 stages. The stages involved are: 1) overall plan, 2) content definition/domain definition and claims statements, 3) test/content specifications, 4) item development, 5) test design and assembly, 6) test production, 7) test administration, 8) scoring test responses, 9) passing/cut scores, and 10) test score reports. The test instruments developed include mathematical literacy indicators according to the OECD, namely formulate, employ, interpret, and reasoning O.e.c.d (2023a) and material on the algebra elements of grade VII, namely algebraic forms and linear equations and inequalities with one variable. The questions developed are in essay form with polytomous scoring.

The subjects of this study were 179 seventh-grade students at a junior high school in Sleman Regency in the 2024/2025 academic year. The sampling technique that was applied for sampling for the purposes of this study was simple random sampling. The instruments used for data collection were content validation sheets and mathematical literacy test sheets. The test instruments were validated by five experts, which include mathematics education lecturers and mathematics teachers. Content validity analysis used the Aiken index.

The programs used to analyze the test results were SPSS, Lisrel 8.80, and Rstudio. SPSS was used to perform exploratory factor analysis (EFA). In EFA analysis, if the Keyser Mayer Oikin (KMO) value, Barlett's Test of Sphericity, and Measures of Adequate Sampling (MSA) meet the requirements, then the next factor analysis can be continued. Lisrel 8.80 was used in CFA (Confirmatory Factor Analysis) to determine the validity and reliability of the construct. The RStudio program was used for IRT analysis with the Partial Credit Model (PCM). PCM is a polytomous scoring model that is an extension of the Rasch model on dichotomous data.

RESULT AND DISCUCSSION

This study purposes to describe the characteristics of test items in a mathematical literacy instrument for seventh-grade algebra using the IRT approach. In developing

instruments, it is important to ensure their quality so that they can produce accurate information. The developed instrument must undergo validation tests to ensure that it can accurately measure what it is supposed to measure (Dewanti et al., 2024). Validity testing was conducted by considering content validity and construct validity. Content validity is used to assess the extent to which the instrument is consistent with the material to be evaluated, while construct validity is used to assess the extent to which the instrument items are able to measure the aspects to be measured based on the underlying theory or conceptual definition (Evendi, 2020). In this study, content validation was conducted by five experts in mathematics education by evaluating the developed items based on content, construction, and language aspects. EFA and CFA were conducted to determine construct validity.

After developing the complete test instrument, content validity testing was conducted by several experts. Items that met the criteria were then used for field test. The students' answers during the field test were then used to test construct validity and construct reliability. Items that met the criteria were then analyzed for item characteristics. In addition, an analysis of person characteristics was also conducted. Items that passed all analyses were retained in the test instrument.

The mathematical literacy instrument for algebraic elements consists of 12 essay questions. Each question measures one mathematical literacy indicator and one material in the seventh-grade algebraic elements. The mathematical literacy indicators used are: 1) formulate (formulate real-world problems mathematically), 2) employ (use mathematical concepts, facts, and procedures to find solutions), 3) interpret (interpret mathematical solutions, results, or conclusions obtained in real-world contexts), and 4) reasoning (provide arguments and evidence to support and explain solutions). The topics included in the algebra elements of seventh grade include algebraic expressions and linear equations and inequalities with one variable.

The content validation results showed that all items met the validity criteria with a value of $V \geq 0.80$, so they could be tested on respondents. Respondent answers were analyzed to determine the characteristics of the items. Before conducting IRT analysis, exploratory factor analysis (EFA) was first performed. The first step was to conduct KMO and Bartlett's tests to evaluate sample adequacy. The analysis results are given in Table 1.

Table 1. Output KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.795
Bartlett's Test of Sphericity	Approx Chi-Square	383.016
	Df	55
	Sig.	.000

Table 1 shows KMO ≥ 0.5, meaning that the total of 179 samples used in the study were adequate. Bartlett's test results show a sig. value of < 0.5, indicating that the data forms a correlation matrix with close relationships between variables.

Based on anti-image correlation, the MSA value obtained was in the interval $0.377 < MSA < 0.86$. This result indicates that one item is not suitable for factor analysis, so that item was excluded. A reanalysis was then conducted, yielding an MSA value within the interval $0.540 < MSA < 0.870$. Therefore, 11 out of 12 items are suitable for factor analysis.

Factor analysis with Lisrel was used to determine construct validity. The determination was made using the standardized loading factor (SLF) value. For a sample size of 150, the minimum SLF value was 0.45 (Hair et al., 2019). The results of the CFA analysis are presented in Figure 1.

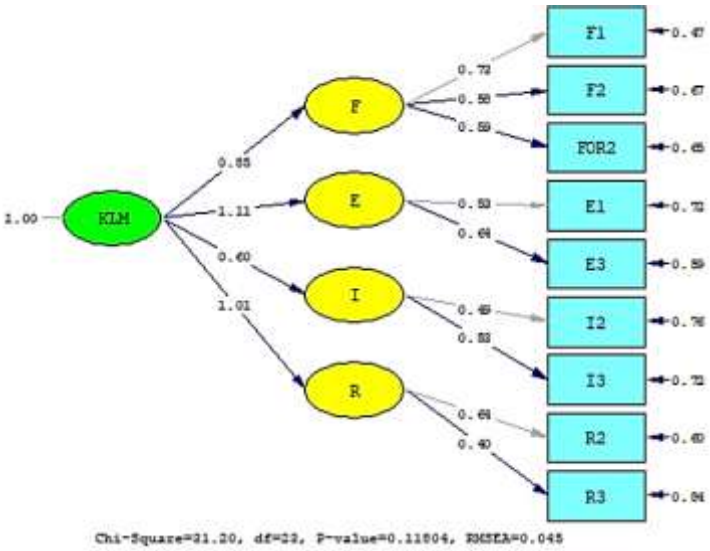

Figure 1. Path Diagram CFA

Figure 1 shows the final results of the CFA. The first CFA removed two items from the model because their SLF values were less than 0.45. In the second CFA, all items met the SLF threshold, resulting in eight items that can be considered construct valid: F1, F2, FOR2, E1, E3, I2, I3, and R2. The SLF value was also used to determine construct reliability (CR). Based on the analysis results, a CR value of 0.81 was obtained. Hair et al. (2019) state that a CR value greater than 0.70 indicates that the mathematical literacy construct is reliable.

The IRT assumptions analyzed consist of unidimensionality, local independence, and parameter invariance. The assumption of unidimensionality is confirmed thorugh the eigenvalues in the scree plot output illustrated in Figure 2.

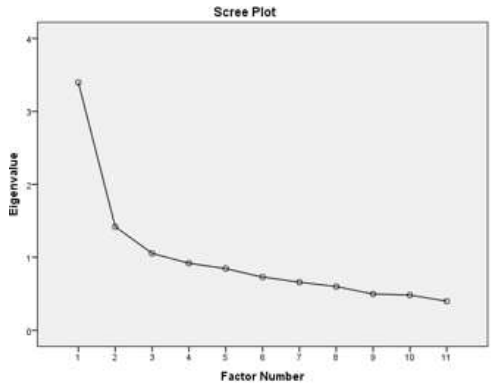

Figure 2. Output Scree Plot

As can be seen in Figure 2, there is only one dominant factor, which then flattens out as the factor number increases. This means that the mathematical literacy test instrument measures one dominant factor. This result is reinforced by the percentage of eigen values in component 1, which is 30.884% of the total. Samritin (2022) states that a variance percentage of more than 20% means that the instrument only contains a single dimension, thus fulfilling the unidimensional assumption When the unidimensional assumption is fulfilled, it also indicates that the local independence assumption has been fulfilled. This aligns with the view of Hambleton et al. (1991) that when the assumption of unidimensionality is fulfilled, the assumption of local independence is also fulfilled.

The third assumption is the item parameter invariance and ability parameter invariance assumption. Item parameter invariance is analyzed by estimating item parameters for odd-numbered and even-numbered respondents. The estimated item parameter is the item difficulty level. The invariance of ability parameters is demonstrated by estimating the ability level on odd and even items. Scatter plots for item parameter invariance and ability invariance are presented in Figure 3.

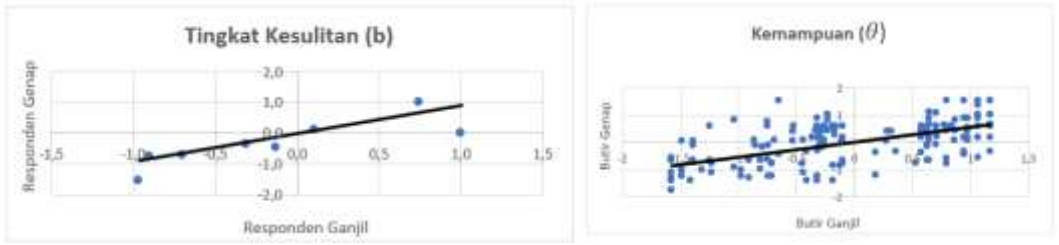

Figure 3. Scatter Plot of Item Parameter Invariance and Ability

Figure 3 shows that each point is relatively close to the line $y = x$. Therefore, the assumptions of item parameter invariance and ability parameter invariance are fulfilled.
The item analysis conducted using IRT is the item difficulty level (b). The item difficulty level parameter (b) refers to the point on the ability scale where an examinee has a 50% chance of answering correctly (Retnawati, 2014). The results are given in Table 2.

Table 2. Estimated Level of Difficulty (b)

Item	Value				
	a	b_1	b_2	b_3	Location
F2	1	0,237	0,202	-0,131	0,102
E1	1	-0,278	-0,802	-1,353	-0,811
E3	1	1,313	-1,544	-1,838	-0,689
R2	1	1,416	0,005	-2,383	-0,320
F1	1	-0,486	3,233	-3,389	-0,214
I2	1	1,696	-4,537	-0,436	-1,092
I3	1	-0,180	0,388	3,433	1,213
FOR2	1	0,184	0,194	1,391	0,839

Description
a = differential power index
b_1 = intersection point of categories 0 and 1
b_2 = intersection point of categories 1 and 2
b_3 = intersection point of categories 2 and 3
Location = average level of difficulty

Based on the table in the location column, it shows that the lowest level of difficulty of the items is -1.092 and the highest is 1.213. This means that all items have a level of difficulty in the range of $-2 < b < 2$, so all items are in the good category (Hambleton & Swaminathan, 1985).

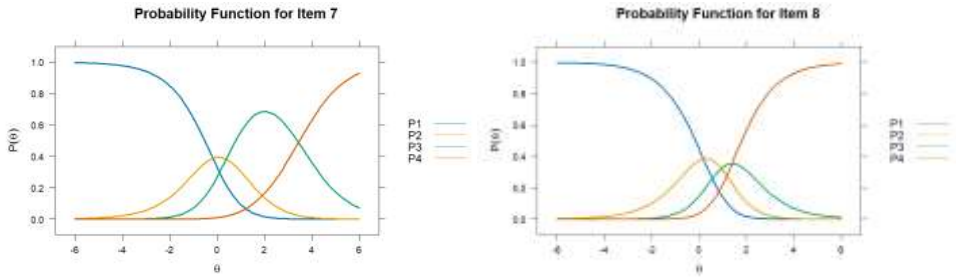

Figure 4. Ideal ICC Plot

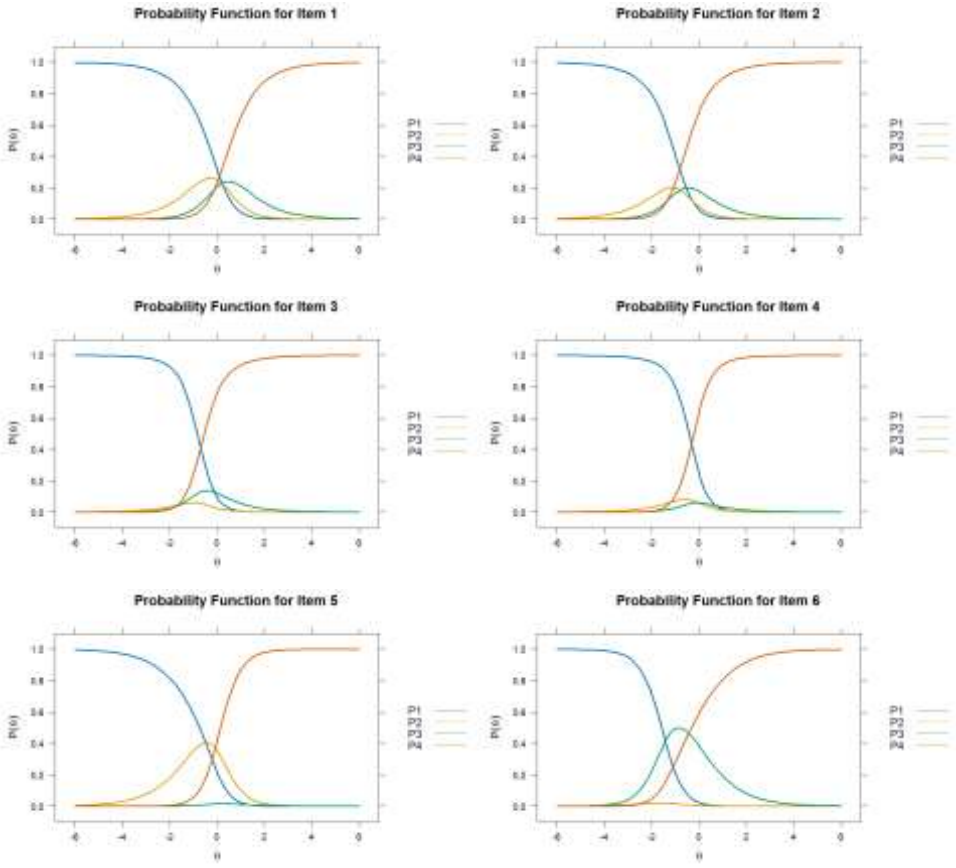

Figure 5. Not Ideal ICC Plot

The results of PCM analysis on IRT also provide Item Characteristic Curves (ICC). The ICC plot illustrates the intersection (threshold) between categories for each item, indicating the level of difficulty required to reach the next category (Mulyani et al., 2017). The ICC curve for a good item has a consistent and ordered pattern, with the crossing point of the n-category and $(n+1)$-category function appearing consistently and without overlap (Dewanti et al., 2024).

An ideal ICC plot is defined by the intersection of curves between response categories that shift increasingly to the right, or in other words, the further to the right the plot, the higher the b value (Dewanti et al., 2024). Generally, thresholds increase sequentially, with each threshold being higher than the previous one. An ideal ICC plot on a item indicates that the item the item can effectively differentiate respondents according to their ability levels. A value of b that does not increase sequentially will cause the ICC plot to be non-ideal because the category functions overlap with each other.

In this mathematical literacy test instrument, there are two items with increasing thresholds from one category to the next, while the other six items have non-sequential thresholds. This is because, according to De Ayala, PCM does not require the steps to solve the items to be sequential, or they must have the same level of difficulty (Dewanti et al., 2020). This results in the thresholds in PCM assessment from one category to the next not always being higher.

This section will discuss one of the ideal ICC plots, which is item I3.

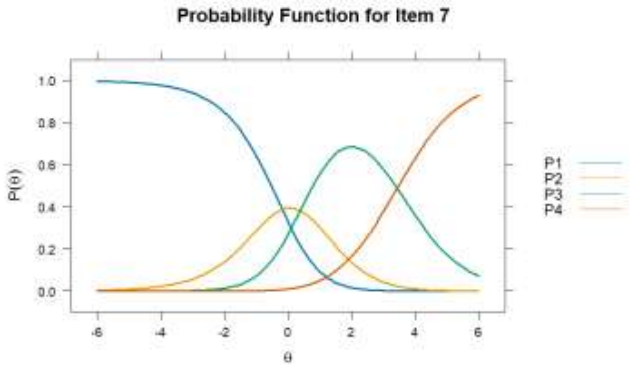

Figure 6. Ordered Threshold Item

Figure 6 shows that the blue and orange line thresholds (categories 0 and 1) are located on the x-axis = -0.180. A score of 1 can be achieved by respondents with a minimum ability of -0.180. A score of 2 can be achieved by respondents with a minimum ability of 0.388. A score of 3 can be achieved by respondents with a minimum ability of 3.43. It can be observed that the threshold values increase sequentially from the lower category to the higher category. Therefore, the ICC plot for item I3 is an ideal plot because it is only natural that higher scores can only be obtained if respondents have higher abilities (Bahar & Retnawati, 2022).

One of the ICC plots that is not ideal is item F1.

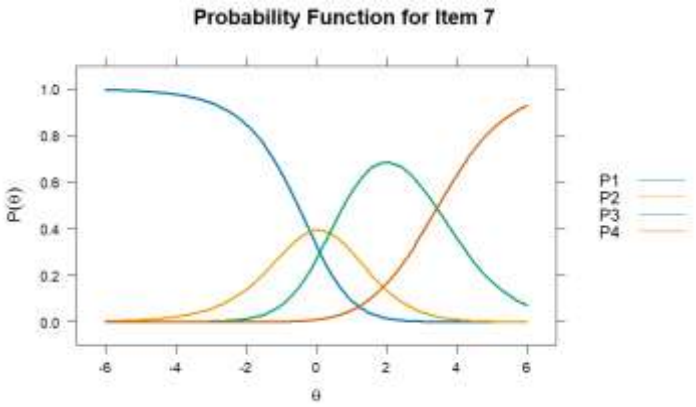

Figure 7. Unordered Threshold Item

The minimum ability of respondents to obtain a score of 1 on item F1 is -0.486. A score of 2 can be achieved by respondents with a minimum ability of 3.233, while respondents with a minimum ability of -3.389 can directly achieve a score of 3. The order of thresholds in this curve is not in the usual order. In Figure 8, it can be seen that the P3 line is very flat. This means that the probability of respondents with low to high ability choosing that category is very small. The P4 line, which represents the highest score category, begins to dominate even at low θ values. Respondents with low ability can directly achieve a score of 3 without passing through score 2. It can be said that category 2 is not functioning optimally.

Master & Wright state that unordered thresholds are not always an indication of problematic items because PCM does not require thesholds to be ordered (Wu & Adams, 2007). When PCM is applied to items whose score categories correspond to sequential "steps" for solving a problem and a threshold disorder occurs, the cause is often that the later steps are easier than the earlier steps (Wu & Adams, 2007). This could be because the frequency of students choosing certain categories is low or because the scoring rubric is unclear in distinguishing between possible answer levels. However, the items in this instrument have been stated to be content valid and therefore suitable for measuring mathematical literacy skills in algebra for seventh grade. Therefore, the items are retained in the instrument with the note that an evaluation of the scoring rubric may be necessary.

Further analysis was conducted to determine the information function and SEM. The test information function is presented in Cartesian coordinates. The x-axis represents the level of respondents's abilities, and the y-axis represents the magnitude of the information function. The values obtained in this analysis are estimates, so their accuracy is probabilistic and subject to measurement error (SEM). The results of the analysis are presented in Figure 8.

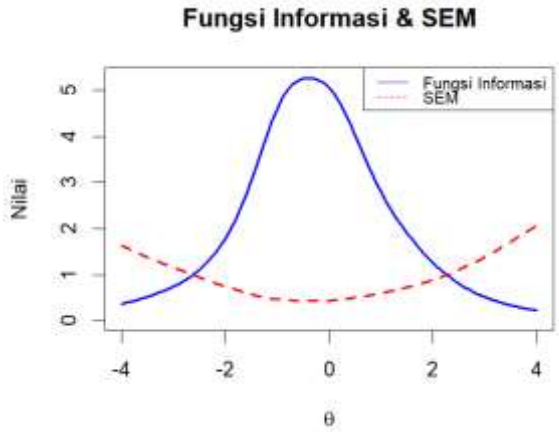

Figure 8. Graph of Function Information and SEM

Figure 8 shows that the top of the test information function is at a value of 5.266 with a measurement error (SEM) of 0.435, which occurs when $\theta = -0.40$. The information function has the largest value when the SEM value is smaller. This is in line with Sumaryanta (2021) opinion, which states that the information function and SEM have an inverse relationship. The intersection of the information function curve and SEM is at -2.6 on the left and 2.25 on the right. It indicates that the test will provide optimal information when administered to respondents with ability levels ranging from -2.6 to 2.25.

Next, item fit analysis was conducted to measure the suitability of respondents' answer patterns to a particular item. The analysis results are shown in Table 3.

Table 3. Item Fit Analysis Results

Question Item	MNSQ	ZSTD	Description
F2	0,827	-1,756	Fit
E1	0,854	-0,854	Fit
E3	0,576	-1,649	Fit
R2	0,692	-1,886	Fit
F1	0,633	-2,630	Not Fit
I2	0,936	-0,326	Fit
I3	0,962	-0,465	Fit
FOR2	0,751	-2,131	Not Fit

An item is categorized according to the IRT model used if $0.5 \leq MNSQ \leq 1.5$ and $-1.9 \leq ZSTD \leq 1.9$ (Linacre, 2002). Based on Table 3, it is shown that items F2, E1, E3, R2, I2, and I3 are fit items. Meanwhile, items F1 and FOR2 are unfit items, meaning they have response patterns that do not match the model's expectations. Unfit items may be caused by defects or problems in the items (Zubairi & Kassim, 2006).

Person fit analysis was based on the same criteria as item fit. Person fit analysis was conducted to identify respondents whose response patterns did not align with the specified model (Reise, 1990). Based on the analysis results from 179 respondents who completed the

test instrument, there were 67 respondents who were not fit or were detected to have different response patterns. There are several potential causes of person misfit, including cheating, careless responding, and lucky guessing (Meijer, 2009).

The cut score for this test instrument was determined using the Angoff method. In this method, several raters set the cut score based on the estimated probability of answering correctly for items grouped according to difficulty level (Retnawati, 2014). The measurements taken by the raters are displayed in Table 4.

Table 4. Measurement Results

Rater	Level of Difficulty		
	Easy	Moderate	Difficult
1	90% 3 items	70% 4 items	30% 1 item
2	90% 4 items	70% 3 items	30% 1 item
3	90% 4 items	70% 2 items	30% 2 items

The cut score for each rater is calculated by adding the product of the percentage and the number of items. If there are multiple raters, the final cut score is the average of the cut scores from each rater. The cut scores from the three raters in this study were 5.8 (if expressed as a score) or 24.16 (if expressed as a value) on a scale of 100. This cut score is used as the passing threshold for respondents. If a respondent scores the same or higher than the cut score, they are assumed to have passed the mathematical literacy test. Out of the 179 respondents in this study, 17 were assumed to have not passed the test.

IRT analyzes students' individual abilities in detail and provides a clear picture of the characteristics of test items. Test instruments analyzed using IRT can be a more accurate evaluation tool for use in mathematics learning. IRT's ability to provide difficulty level parameters allows teachers to identify which items work well and which may need to be revised or removed (Fitraynsyah et al., 2024). In addition, IRT models the probability of correct responses for each item based on the latent properties being measured, such as students' knowledge levels (Fitraynsyah et al., 2024). Based on this information, teachers can identify students' learning difficulties and design appropriate learning strategies.

Mathematical literacy is an individual's ability to recognize the role of mathematics and use it to solve everyday problems. The mathematical literacy instrument developed in this study uses contextual problems that are relevant to students' lives. By presenting relevant contextual problems, students not only learn to understand mathematical concepts abstractly, but also see how these concepts are used in real situations (Hidayatulloh et al., 2025). This is in line with constructivist theory, which states that knowledge is constructed by individuals through experience and interaction with the environment (Azzahra et al., 2025). The use of contextual problems in mathematical literacy instruments allows students to relate the mathematical concepts they have learned to the everyday problems they face.

In solving contextual problems, especially in algebra, each student's response will certainly differ. This diversity in student responses can occur due to differences in the thought processes or knowledge levels of each student and the level of difficulty of the questions. Therefore, item analysis using IRT is used to describe the characteristics of the items and the relationship between student ability and the probability of answering a question correctly more accurately.

CONCLUSION

The instrument for testing mathematical literacy in algebra for seventh graders that was developed has good quality according to the validity of content, construct validity, and construct reliability. The characteristics of the items in the mathematical literacy test instrument for the algebra element of grade VII are as follows: (1) all items have a difficulty level within the range $-2 < b < 2$, so the items are in the good category; (2) the discrimination power of all items is one; (3) the item characteristic curve (ICC) shows that there are 2 items with ideal ICC and 6 items with non-ideal ICC; (4) The information function and SEM indicate that the test instrument provides the best information when administered to students with abilities slightly below average; (5) Item fit results show that 6 items fit the model, while person fit results show that 112 students fit the model.

This mathematical literacy test instrument has good quality and characteristics of test items, so it is used to measure students' mathematical literacy skills. Information about students' abilities can be used as a basis for designing learning activities that support mathematical literacy. This study can be used as a guideline in developing test instruments to measure students' mathematical literacy skills. Future researchers are expected to make more detailed scoring rubrics in order to better distinguish students' skill levels. In addition, further research could try to analyze using models other than PCM.

REFERENCES

Apriatni, S., Yuhana, Y., & Sukirwan. (2022). Pengembangan Instrumen Literasi Numerasi Materi Trigonometri Kelas X SMA. *EDU-MAT: Jurnal Pendidikan Matematika*, *2759*(2), 185–198. https://doi.org/10.20527/edumat.v10i2.13720

Ardiansari, L. (2018). Pra-aljabar: Langkah Baru Mengajar Aljabar Awal (Penerapan Didactical Design Research. *Proximal: Jurnal Penelitian Matematika Dan Pendidikan Matematika*, *1*(1), 32–44. https://doi.org/https://e-journal.my.id/proximal/article/view/182

Azzahra, N. T., Ali, S. N. L., & Bakar, M. Y. A. (2025). Teori Konstruktivisme dalam Dunia Pembelajaran. *Jurnal Ilmiah Research Student*, *2*(2), 64–75. https://doi.org/10.61722/jirs.v2i2.4762

Bahar, R., & Retnawati, H. (2022). Analisis Karakteristik Soal Kemampuan Koneksi Matematika Penskoran Politomus. *Jurnal Tarbiyah*, *29*(2), 195–211. https://doi.org/10.30829/tar.v29i2.1650

Cendana, V., Syamsuri, & Pujiastuti, H. (2024). Pengembangan Instrumen Literasi Matematis Model PISA dengan Konteks Batik Krakatoa Cilegon untuk Siswa SMP. *Himpunan: Jurnal Ilmiah Mahasiswa Pendidikan Matematika*, *4*(1), 29–40. https://doi.org/https://jim.unindra.ac.id/index.php/himpunan/article/view/11188

Dewanti, S. S., Ayriza, Y., & Setiawati, F. A. (2020). The Application of Item Response Theory for Development of a Students' Attitude Scale Toward Mathematics. *New Educational Review*, *60*, 108–123. https://doi.org/10.15804/tner.2020.60.2.09

Dewanti, S. S., Hadi, S., Nu'man, M., & Ibrahim. (2021). The Application of Item Response Theory in Analysis of Characteristics of Mathematical Literacy Test Items. *Ilkogretim Online - Elementary Education*, *20*(1), 1226–1237. https://doi.org/10.17051/ilkonline.2021.01.119

Dewanti, S. S., Izzah, J. N., Kiranasari, S. P., & Sholihin, K. F. (2024). Utilizing Item Response Theory for the Analysis of Self-regulated Learning Scale in Mathematics. *Jurnal Elemen*, *10*(June), 614–629. https://doi.org/10.29408/jel.v10i3.26618

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates.

Evendi, E. (2020). Evaluasi Pembelajaran Matematika.

Farida, R. N., Qohar, A., & Rahardjo, S. (2021). Analisis Kemampuan Literasi Matematis Siswa SMA Kelas X dalam Menyelesaikan Soal Tipe PISA Konten Change and Relationship. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *05*(03), 2802–2815. https://doi.org/https://j-cup.org/index.php/cendekia/article/view/972

Fitraynsyah, M. A., Hilmiyati, F., & Habudin. (2024). Peran Tingkat Kesukaran dan Daya Pembeda dalam Analisis Butir Tes : Kajian Literatur untuk Pendidikan Menengah. *JREP: Jurnal Riset Dan Evaluasi Pendidikan*, *1*(4), 252–262. https://doi.org/10.51574/jrep.v1i4.2250

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis (Eighth*. Cengage Learning EMEA.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory*. Kluwer Inc.

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

Hidayatulloh, D. A., Agustin, D. R., & Malik, F. A. (2025). Menumbuhkan Literasi Matematis Siswa dalam Pembelajaran Melalui Lensa Konstruktivistime. *Wacana Akademika: Majalah Ilmiah Kependidikan Volume*, *9*(1), 99–105. https://doi.org/10.30738/wacanaakademika.v9i1.19786

Linacre, J. M. (2002). What Do Infit and Outfit, Mean-square and Standardized Mean? *Rasch Measurement Transactions*, *16*(2), 878.

Mardapi, D. (2017). *Pengukuran, Penilaian, dan Evaluasi Pendidikan* (2nd ed.). Parama Publishing.

Meijer, R. R. (2009). Person-Fit Research : An Introduction. *Applied Measurement in Education*, *9*(1), 37–41. https://doi.org/10.1207/s15324818ame0901

Mulyani, S., Efendi, R., & Ramalis, T. R. (2017). Karakterisasi Tes Keterampilan Pemecahan Masalah Fisika berdasarkan Teori Respons Butir.

Nursakiah, N., Arriah, F., & Dharma, S. (2022). Developing Mathematical Literacy Test with Context of Bugis-Makassar Local Wisdom for Junior High School Students. *Jurnal Elemen*, *8*(1), 16–28. https://doi.org/10.29408/jel.v8i1.4049

O.e.c.d. (2023a). *PISA 2022 Assessment and Analytical Framework. In OECD (Organisation for Economic Co-operation and Development) Publishing*. OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en

O.e.c.d. (2023b). Pisa 2022 Results (Volume I): The State of Learning and Equity in Education. In *OECD Publishing* (Vol. I). OECD Publishing. https://doi.org/10.1787/53f23881-en

Ojose, B. (2011). Mathematics Literacy: Are We Able To Put The Mathematics We Learm Into Everyday Use. *Journal Of Mathematics Education*, *4*(1), 89–100. https://educationforatoz.com/images/Bobby_Ojose.pdf

Reise, S. P. (1990). A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137.

Retnawati, H. (2014). *Teori Respon Butir dan Penerapannya*. Nuha Medika.

Samritin. (2022). Identifikasi Muatan Differential Item Functioning Pada Data Ujian Nasional Matematika. *Journal on Education*, *04*(04), 1675–1684. https://doi.org/10.31004/joe.v4i4.2508

Santoso, A. (2018). Karakteristik Butir Tes Pengantar Statistika Sosial berdasarkan Teori Respon Butir. *Jurnal Pendidikan Matematika dan Sains*, *VI*(2), 1–11. https://doi.org/10.21831/jpms.v4i1.10111

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory vs Item Response Theory? Didaktika. *Jurnal Pendidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296

Selan, M., Daniel, F., & Babys, U. (2020). Analisis Kemampuan Literasi Matematis Siswa dalam Menyelesaikan Soal PISA Konten Change and Relationship. *AKSIOMA: Jurnal Matematika Dan Pendidikan Matematika*, *11*(2), 335–345. https://doi.org/10.26877/aks.v11i2.6256

Siregar, N. H., Remiswal, & Khadijah. (2024). Analisis Butir Soal Ujian Tengah Semester pada Mata Pelajaran Pendidikan Agama Islam. *Urwatul Wutsqo: Jurnal Studi Kependidikan Dan Keislaman*, *13*(2), 179–189. https://doi.org/10.54437/juw

Sudaryono. (2012). *Dasar-Dasar Evaluasi Pembelajaran*. Graha Ilmu.

Sumaryanta. (2021). *Teori Tes Klasik & Teori Respon Butir (Konsep & Contoh Penerapannya*. CV Confident.

Wibowo, A. A., Rif'at, M., & Yani, A. (2020). Pengembangan Instrumen Tes Untuk Mengukur Kemampuan Literasi Matematis Siswa SMP. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *9*(7). https://doi.org/10.26418/jppk.v9i7.41316

Wu, M., & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. In *Educational Measurement Solutions*.

Zahrah, M. (2024). Penelitian Literasi Matematis di Sekolah: Pengertian dan Kesulitan-Kesulitan Siswa. *Jurnal Riset Pendidikan Matematika Jakarta*, *6*(1), 27–36. https://doi.org/10.21009/jrpmj.v6i1.29024

Zubairi, A. M., & Kassim, N. L. A. (2006). Classical and Rasch Analyses of Dichotomously Scored Reading Comprehension Test Items. *Malaysian Journal of ELT Research*, *2*(March), 1–20. https://doi.org/https://meltajournals.com/index.php/majer/article/view/663