

Kalibrasi Instrumen Literasi Matematika Siswa Menggunakan IRT dan Aplikasinya untuk Estimasi Skor

Hari Purnomo Susanto¹, Heri Retnawati²

¹Sekolah Tinggi Keguruan dan Ilmu Pendidikan PGRI Pacitan

²Penelitian dan Evaluasi Pendidikan Universitas Negeri Yogyakarta

Email: haripsusanto@stkippacitan.ac.id¹, heri_retnawati@uny.ac.id²

Abstrak

Literasi matematika merupakan salah satu komponen yang menjadi perhatian pada Assesmen Kompetensi Minimum (AKM). Kebijakan ini sebagai upaya Kemendikbud untuk mengatasi rendahnya kemampuan literasi siswa, dan faktor-faktor penyebabnya. *Multistage Adaptive Test* (MSAT) digunakan sebagai metode penilaian pada AKM. MSAT dikembangkan dengan konsep IRT. Tujuan dari artikel ini yaitu mengaplikasikan *Item Respons Theory* (IRT) untuk kalibrasi instrumen literasi matematika, dan memanfaatkan parameter butir hasil kalibrasi untuk estimasi skor kemampuan literasi matematika. Banyak butir pada instrumen yang digunakan yaitu 15 butir pilihan ganda. Responden yang digunakan sebanyak 65 siswa SMP kelas 8. Setelah melalui proses validitas konstruk menggunakan *Confirmatory Factor Analysis* (CFA) 2 butir tidak memenuhi, karena memiliki faktor loading yang rendah dan 13 butir digunakan untuk proses kalibrasi. Proses kalibrasi butir dan penskoran dilakukan dengan program R *package mirt*. Hasil kalibrasi menunjukkan bahwa instrumen cocok dengan model *Rasch*. Semua asumsi IRT unidimensi terbukti terpenuhi, sehingga tidak ada pelanggaran dalam mengestimasi parameter butir. Parameter butir berupa tingkat kesulitan dimana terdapat satu butir tidak fit yaitu butir ke-3. Perpotongan fungsi informasi dan standar *error* menunjukkan instrumen akan memberikan informasi yang akurat, jika digunakan siswa dengan kemampuan -2.965 sampai 1.085 . Parameter butir yang dihasilkan dapat digunakan untuk estimasi skor kemampuan literasi matematika konten aljabar.

Kata Kunci: CFA, IRT, kalibrasi, konten Aljabar, literasi matematika, penykoran IRT

Calibration of Students' Mathematical Literacy Instrumens using IRT and Its Applications for Score

Abstract

Mathematical literacy is one of the components that is of concern to the Minimum Competency Assessment (AKM). This policy is an effort by the Ministry of Education and Culture to overcome the low literacy skills of students, and the faktors that cause it. Multistage Adaptive Test (MSAT) is used as an assessment method in AKM. The MSAT was developed with the IRT concept. The purpose of this article is to apply Item Respons Theory (IRT) to calibrate mathematical literacy instrumens and utilize the calibration result item parameters to estimate scores for mathematical literacy skills. Many items in the instrumen used are 15 multiple choice items. Respondents were used as many as 66 grade 8 junior high school students. After going through the construct validation process using Confirmatory Factor Analysis (CFA), 2 items did not meet, because they had a low loading factor, and 13 items were used for the calibration process. The grain calibration and scoring processes were carried out using the R package Mirt program. The calibration results show that the instrument matches the Rasch model. All assumptions of unidimensional IRT are proven to be met, so there is no violation in estimating item parameters. The item parameter is the level of difficulty where there is one item that does not fit, namely item 3. The intersection of the information function and the standard error shows that the instrument will provide accurate information if it is used by students with abilities of -2.965 to 1.085 . The resulting item parameters can be used to estimate the algebra content math literacy score.

Keywords: algebra content; CFA; IRT calibration; IRT scoring; mathematical literacy

INTRODUCTION

Mathematical literacy is defined as students' ability to formulate, use, and interpret mathematics in various contexts. This includes mathematical reasoning and the use of mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena (OECD, 2013). The Department of Higher Education and Training (DHET) (Vale et al., 2013) views mathematical literacy as an individual attribute that involves managing situations and solving problems in everyday life, work, and society by engaging with mathematical concepts. Mathematical literacy can describe how an individual applies mathematics in daily life (Hasanah & Hakim, 2022; Yuberta et al., 2020).

The track record related to mathematical literacy skills of junior high school students in Indonesia is very low (Hasanah & Hakim, 2022; Hertiantito, 2016; Larasaty et al., 2018; Yuberta et al., 2020). There are many factors that contribute to the low levels of mathematical literacy achievement among Indonesian students according to the OECD report. The Indonesian curriculum does not sufficiently support the development of mathematical literacy in the classroom (Masjaya & Wardono, 2018). Students' lack of reference to PISA-type problems in mathematics learning (Mansur, 2018), makes PISA-type problems non-routine for students. The role and function of teachers in mathematics learning, especially in how they deliver the lesson material, has remained unchanged (Masjaya & Wardono, 2018), and teachers tend to use conventional methods in teaching.

To address these problems, in 2021, the Ministry of Education and Culture (Kemendikbud) released a policy on Minimum Competency Assessment (AKM). One of the components of students' learning outcomes measured is mathematical literacy (numeracy). The AKM assessment system adopts the PISA assessment system, which uses MSAT (Kemendikbud, 2021). This assessment system is very different from the previous National Examination (UN) assessment system, which used a Computer-Based Test (CBT) system. In the MSAT system, there is a bank of items used as the main object in measuring students' abilities. This item bank consists of test items that have been calibrated using the IRT concept (Rotou et al., 2007).

The concept of Item Response Theory (IRT) is still rarely used in educational practice in Indonesia. In practice, the most familiar concept used is the Classical Test Theory (CTT) (Dewanti et al., 2021). The IRT concept must be understood by teachers in relation to standardized tests that can be used on a large scale (national) rather than locally. The IRT scoring system is different from traditional scoring, as it is based on the item parameters (Brown, 2018). These item parameters are processed by MSAT to measure the students' abilities while taking the test.

A good test instrument should have the property of invariance. Invariance of parameters is a fundamental measurement property that cannot be ignored (Rupp & Zumbo, 2006). This property indicates the instrument's ability to function as a measuring tool with the same characteristics when used to measure different groups or over time (Finch, 2014). The instrument's characteristics that meet the property of invariance are needed in AKM. According to the purpose of AKM, which is to evaluate the system, it requires large-scale assessments whose results can be used to compare scores between groups of students, schools, regions, and over time in an objective manner (Kemendikbud, 2019).

Invariance can be empirically proven from the perspective of item and participant. Two conditions of invariance that must be met are (1) item parameters should not be influenced by participants' abilities, and (2) participants' abilities should not be influenced by item parameters (Retnawati, 2014). Both of these conditions cannot be met by the CTT model (Dewanti et al., 2021). Invariance property becomes an assumption that must be met when using IRT for instrument calibration (Nguyen et al., 2014; Retnawati, 2014). The proof of parameter invariance can be done by detecting Differential Item Functioning (DIF) (Nguyen et al., 2014; Retnawati, 2014).

The article applies IRT for calibrating a mathematics literacy instrument and utilizes the resulting item parameters to estimate mathematics literacy ability scores. The calibration results are further utilized for determining the scoring of students' mathematics literacy abilities in Algebra content.

METHOD

This study uses a descriptive method with a quantitative approach, aiming to empirically analyze the characteristics of an IRT-based mathematics literacy test. The mathematics literacy instrument adopted existing questions but with characteristics in specific areas of mathematics literacy. The instrument used focuses only on algebra content, and the questions used can be seen in Table 1. The instrument was tested on 65 respondents of 8th grade junior high school students. The data was used to prove the construct validity, reliability, and item calibration in Table 1. The calibration results were used as the basis for estimating student literacy scores. Before conducting IRT-based instrument calibration, the instrument must undergo validity and reliability testing. The validity used in this article is construct validity, proven through CFA analysis. The CFA procedure used in this study includes (1) calculating sample adequacy, (2) testing for multicollinearity (Tabachnick & Fidell, 2014), (3) conducting CFA order-1 analysis, and (4) if CFA order-1 is not met, then conducting CFA order-2 (Retnawati, 2016). An instrument is considered to have good construct if the model fits the empirical data, resulting in a p-value $\chi^2 > 0.05$, RMSEA < 0.08 , and CFI > 0.9 (Hair et al., 2018), and using factor loading not less than 0.4 (Hair et al., 2018; Retnawati, 2016). Furthermore, the reliability testing used in this study is Construct Reliability (CR), based on the factor loading and error of each indicator/item. The formula for calculating CR can be seen in Formula 1.

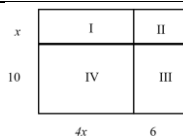
$$CR = \frac{(\sum_1^n L)^2}{(\sum_1^n L)^2 + (\sum_1^n E)} \quad (1)$$

In Formula 1, L and E represent the factor loading and error values of each indicator/item, respectively. n indicates the number of items that have a factor loading of at least 0.4. A CR value of ≥ 0.7 (Hair et al., 2018) is considered to have acceptable reliability. The construct validity calculation with CFA was performed using the R package hpsCFA (Susanto, 2022), and CR was calculated using Excel.

Item calibration was carried out using the IRT concept. The mathematical literacy instrument in this article was calibrated using the following procedures: (1) Determine the model fit with the M2 method. (2) Determine the best model based on RMSEA M2 (Chalmers, 2012; Maydeu-Olivares, 2013; Paek & Cole, 2019). (3) Test the unidimensionality assumption. (4) Test the local independence assumption. (5) Verify item invariance (Nguyen et al., 2014; Retnawati, 2014). (6) Determine item fit. (7) Interpret the information function. All of these procedures were analyzed using the R package mirt (Chalmers, 2012; R Core Team, 2022).

After item calibration, ability score estimation can be performed. There are several methods for estimating the ability of test-takers in the IRT concept. (1) Maximum Likelihood Estimator (MLE), Weighted Likelihood Estimator (WLE), and (2) Bayesian Estimators (Magis & Barrada, 2017). Bayesian Estimators consist of two methods, namely Maximum A Posteriori (MAP), or commonly called Bayesian Modal, and Expected A Posteriori (EAP) (Magis et al., 2017). In addition to these methods, Chalmers (2012) added several other estimation methods, namely EAPsum, Plausible, and Classify. The ability score estimation for algebraic content literacy in this article used the EAP estimation method and was calculated using the R package mirt. The score limit for algebraic content literacy used in this article was between -3 and 3.

Table 1. Algebra Literacy Test Items

Nomer	Test Items	Nomor
3	<p>A window design has a rectangular shape with a length of $(4x + 6)$ cm and a width of $(x + 10)$ cm divided into 4 parts as shown in the figure. Which of the following is an incorrect equation for the area of the window?</p> 	Algebraic Forms
5	A rectangular-shaped rice field is 5 m longer than its width. If the width is x meters, then the area of the rice field is...	Algebraic Forms
8	Given the area of a rectangular-shaped tile is $m^2 + 5m - 50$ cm ² , with a length of $m + 10$ cm, what is the width of the tile?	Algebraic Forms
15	A hairdresser charges Rp15,000 per customer. On average, the hairdresser spends Rp25,000 per day. If the hairdresser's income is represented by y and the number of customers is represented by x , the appropriate equation to represent the hairdresser's daily income is..	Algebraic Forms
1	The difference between the amount of money owned by a younger sibling and an older sibling is Rp10,000. Twice the amount of money owned by the older sibling plus the amount of money owned by the younger sibling equals Rp50,000. What is the total amount of money owned by both siblings?	SPLDV
2	A student buys 5 books and 3 pencils for Rp22,500 at the same store. Another student buys 6 books and 3 pencils for Rp25,500. If you buy a book and 2 pencils at the store, how much do you have to pay?	SPLDV
10	A is four times older than B. If five years from now, A's age will be three times B's age, then the current ages of A and B are	SPLDV
11	A parking lot can accommodate 96 units of motorcycles and cars. If the total number of wheels is 256, which of the following statements is true?	SPLDV
12	The sum of the scores of two volleyball teams in the 23rd minute is 35, and the difference between the scores is 5. What is the product of the scores of the two teams?	SPLDV
13	A grandfather has 6 fewer cows than ducks. If the total number of feet of cows and ducks is 36, how many ducks does the grandfather have?	SPLDV
14	A younger sibling is 3 years younger than the older sibling. If their total age is 19 years, what is the ratio of the older sibling's age to the younger sibling's age in 4 years?	SPLDV
4	A mother has a certain amount of money. She spends one-quarter of it at the market and one-third of the remaining amount on transportation. If she has Rp20,000 left, how much money did she have at first?	Proportion
6	A car needs 11 liters of fuel to travel 121 km. On another occasion, the car travels 71.5 km with 10 liters of fuel in the tank. How many liters of fuel are left in the car?	Proportion
7	Ten rice planters can plant rice in one paddy field in 6 hours. The owner of the field wants the rice planting to be completed in 4 hours. How many rice planters should the owner hire?	Proportion
9	In September, an accessories key seller is able to sell $\frac{3}{5}$ of the total accessories they had in August. What is the ratio of the sold accessories to the unsold accessories?	Proportion

RESULTS

This section will discuss the validation of construct validity, instrument reliability, and subsequent item calibration and its application in estimating mathematical literacy scores. Based on the construct in Table 1, the construct validity was tested using the CFA analysis procedure as previously described

Determining sample adequacy

The KMO value was calculated and obtained a value of $0.66 > 0.5$, indicating that the sample used was adequate for factor analysis

Testing the assumption of multicollinearity among test items

Multicollinearity assumptions were tested by examining the correlation level among items. The results can be seen in Figure 1. To read Figure 1, it is sufficient to observe the numbers printed in bold, which should be less than 0.7 (Hair et al., 2018), and ignore the main diagonal.

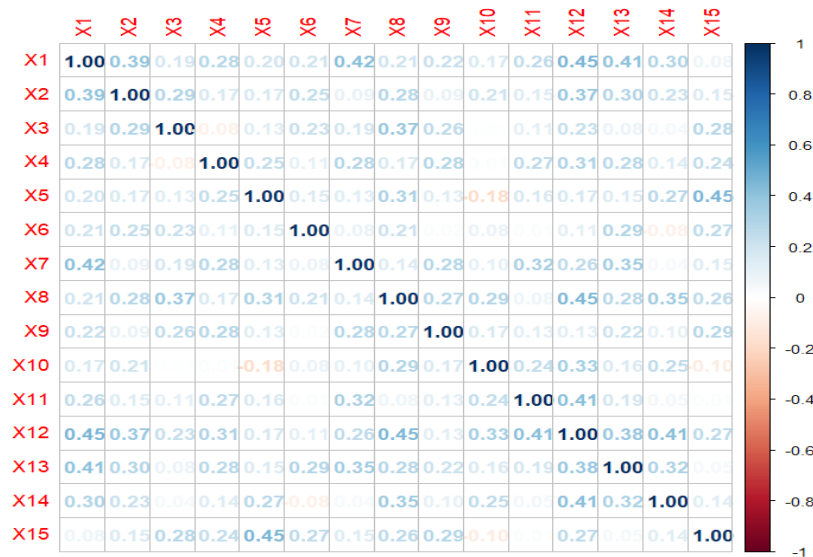


Figure 1. Correlation Between Items

In Figure 1, it is clear that there is no high correlation among the items or no correlation coefficient values are greater than 0.7. Thus, there is no multicollinearity issue among the instrument items (Hair et al., 2018).

Construct Validity Proof with CFA Order 1.

The result of the CFA analysis using the hpsCFA package obtained a model fit after removing items 6 and 10. The result can be seen in Figure 2. The instrument construct model in Figure 2 was analyzed with CFA order 1. The result of the CFA order-1 analysis showed a model fit with a P-value of 0.26, CFI of 0.949, and RMSEA of 0.041. The construct model in Figure 2 has demonstrated a good fit with the empirical data used, and each item has a factor loading of not less than 0.4. However, in this case, multicollinearity occurred in the latent constructs of SPLDV and Proportion. This is indicated by the correlation coefficient value between SPLDV and Proportion of $0.71 > 0.7$ (Hair et al., 2018). This result explains that the CFA order 1 model cannot be used, so it must be continued with CFA order-2.

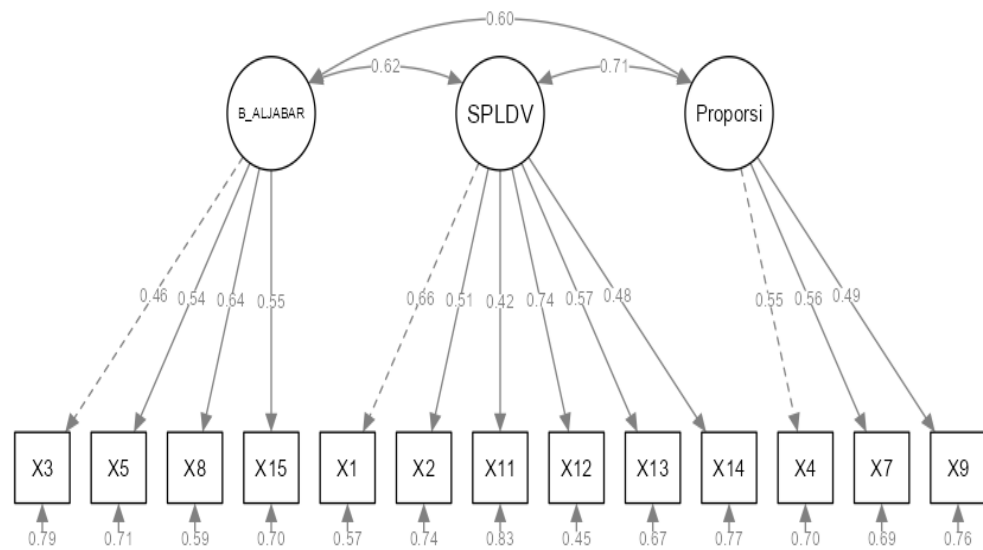


Figure 2. Instrument Construct Model with Order-1 CFA

Validating the Construct with CFA Order 2.

In the CFA Order 2 analysis, a model fit was achieved by removing items 6 and 10. The P-value, CFI, and RMSEA values were the same as those in CFA Order 1. The instrument construct model can be seen in Figure 3. The results in Figure 3 indicate that each item's factor loading has a value greater than or equal to 0.4. This result indicates that the mathematical literacy instrument construct can be used with a valid combination of items, except for items 6 and 10. These two items were excluded from the figure and the item calibration process using IRT.

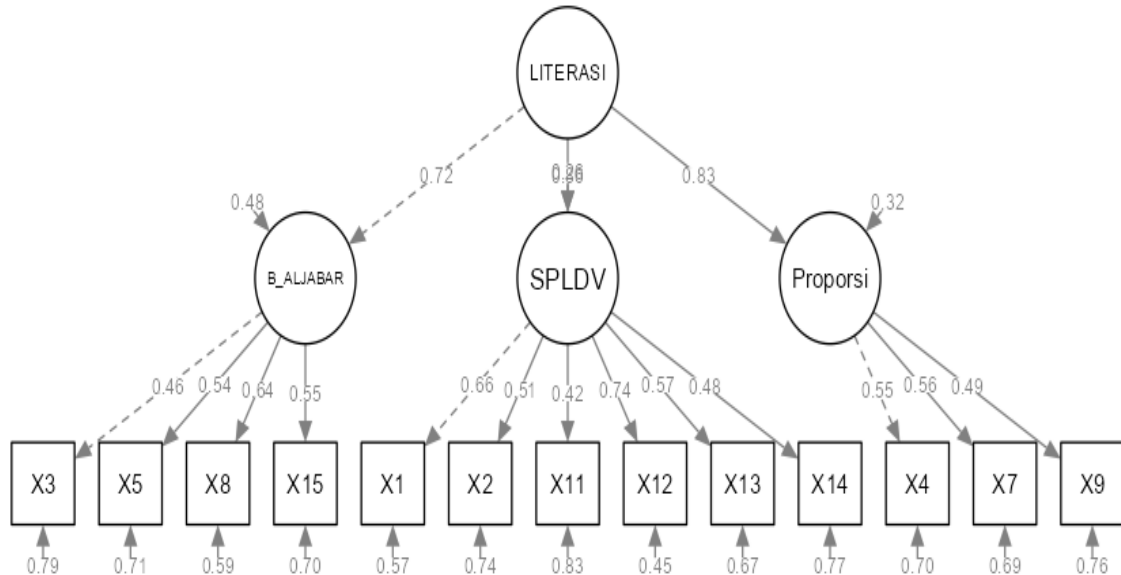


Figure 3. Instrument Construct Model with CFA Order-2

Construct Reliability

The construct reliability was calculated using formula 1 based on the factor loadings and errors obtained from Figure 3, as shown in Table 2. The construct reliability was found to be $CR = 0.8335 > 0.7$. This result indicates that the instrument has high reliability or consistency when used to measure students' mathematical literacy.

Table 2. Loading and Error Factors for Each Item

Item	Factor Loading	Error
3	0.46	0.79
5	0.54	0.71
8	0.64	0.59
15	0.55	0.7
1	0.66	0.57
2	0.51	0.74
11	0.42	0.83
12	0.74	0.45
13	0.57	0.67
14	0.48	0.77
4	0.55	0.7
7	0.56	0.69
9	0.49	0.76

Based on the results of the construct validity, 13 items were found to be valid for use in the calibration process, while items 6 and 10 were excluded. These two items can still be used in the calibration process, but only after they have been improved and reviewed by experts. However, in this article, it was decided to use the 13 valid items for the calibration process. The calibration of the 13 items was performed using the R package mirt, and the results are presented below.

Model Fit Determination

Table 3 shows that each IRT model used has a p-value > 0.01, RMSEA < 0.8, and CFI > 0.9 (Maydeu-Olivares, 2013, 2014). These values indicate that each dichotomous model in Table 3 is a good fit for use.

Table 3. Model Fit

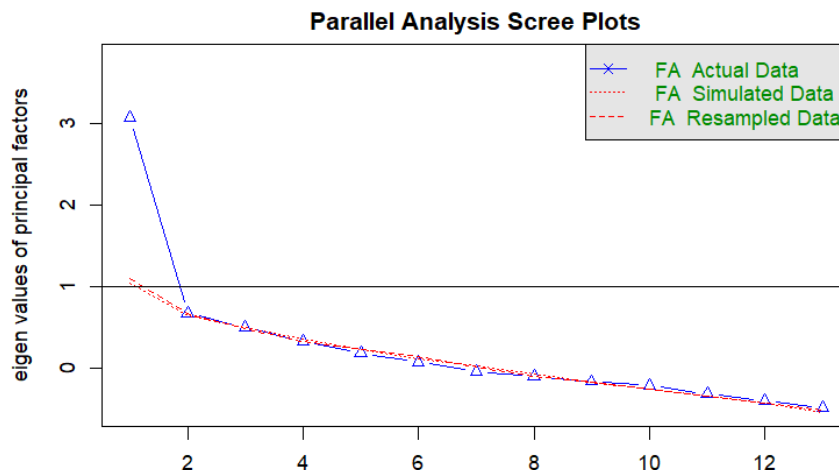
Model	M2	df	p-Value	RMSEA	CFI
<i>Rasch</i>	78.49	77	0.43	0.02	0.99
<i>2PL</i>	68.65	65	0.35	0.03	0.99
<i>3PL</i>	50.09	52	0.55	0.00	1.00

Determining the Best IRT Model

The best dichotomous model was chosen based on the recommendation from (Maydeu-Olivares, 2014; Paek & Cole, 2019) using the smallest value of RMSEA M2. Table 3 shows that the smallest RMSEA value is obtained by the 3PL model, but this model requires a large amount of data to be used. Based on this information, the Rasch model was chosen considering that only one parameter, namely the difficulty parameter, is estimated and a small sample size of 65 students was used in this study.

Proving Unidimensionality

The article proves unidimensionality using a scree plot from parallel analysis factor analysis method. The scree plot can be seen in Figure 4. The plot shows that the instrument has a single dimension, visually indicated by only one steep slope. Many steep slopes indicate multiple dimensions (Retnawati, 2014). This result is supported by the ratio of the first eigenvalue to the total eigenvalue, which is 76.8%. This result indicates that the first factor contributes more than 20% of the variance (Hambleton et al., 1991; Retnawati, 2014).



Gambar 4. Scree Plot Unidimensi

Proving Local Independence

The absence of local dependence in this study was demonstrated using the Q_3 method. Local dependence occurs if there are absolute Q_3 values greater than 0.2236 (Paek & Cole, 2019). The results of the Q_3 calculations can be seen in Table 4 columns 1-3.

Table 4. Q_3 Value to Detect Local Dependency

No	Q_3	$ Q_3 $	Item Pair	Pair
1	-0,376	0,376	(x_3, x_4)	There is no sub element content relationship
2	-0,317	0,317	(x_1, x_{15})	There is no sub element content relationship
3	-0,316	0,316	(x_9, x_{12})	There is no sub element content relationship
4	-0,297	0,297	(x_{13}, x_{15})	There is no sub element content relationship
5	-0,294	0,294	(x_7, x_{14})	There is no sub element content relationship
6	0,288	0,288	(x_5, x_{15})	There is no relation to the settlement procedure
7	-0,261	0,261	(x_8, x_{11})	There is no sub element content relationship
8	-0,248	0,248	(x_5, x_{12})	There is no sub element content relationship
9	-0,238	0,238	(x_{11}, x_{15})	There is no sub element content relationship
10	-0,237	0,237	(x_7, x_8)	There is no sub element content relationship
11	-0,235	0,235	(x_3, x_{14})	There is no sub element content relationship
12	-0,232	0,232	(x_1, x_8)	There is no sub element content relationship
13	-0,232	0,232	(x_2, x_7)	There is no sub element content relationship
14	-0,224	0,224	(x_2, x_9)	There is no sub element content relationship

The calculation results of Q3 in Table 4 indicate the occurrence of local dependence in the mathematical literacy instrument. The second column shows that the $|Q3|$ value is greater than 0.2236, and the pairs of items that are indicated to cause local dependence can be seen in column 4. Furthermore, column 5 shows that there is no relationship between the sub-element of algebra and the solution procedure in each pair of items in column 3. The relationship between sub-elements or solution procedures can be seen in Table 1. Since there is no relationship between sub-elements in algebra in Table 4 and there is no relationship between solution procedures in pairs of items 5 and 15, local independence can be ignored. Regarding the case of local independence, it will be discussed in depth in the discussion section.

Prove The Invariance Property Of The Instrument

To prove the invariance property of the instrument, this article conducted two stages as follows. (1) To prove the invariance of the item parameters of the mathematical literacy instrument, the data was split into two groups based on odd and even patterns. Subsequently, item parameter estimation was conducted based on the Rasch model (the model fit used) for both groups. The obtained parameters were correlated and the results are presented in Table 5. (2) To prove the invariance of the ability parameters, the instrument was divided into two sets of tests based on odd and even patterns. Next, the item parameters of the two tests were estimated using the Rasch model, and the results were used to estimate the ability of all students for both tests. The results of the analysis can be seen in Table 5.

Table 5. Correlation Coefficient of Detailed Parameters and Ability to Invariance Proof.

Parameter	Correlation	p_Value
T_Kesulitan	0.709	0.007
Kemampuan	0.669	0.000

Table 5 shows that the correlation between item parameters estimated from two groups of respondents is strong, indicating that the item parameters meet the criterion of parameter invariance. Furthermore, the correlation between ability estimates from two different tests is moderate, suggesting that the item parameters do not affect the estimation of ability parameters. Overall, there is no violation of the invariance of the mathematics literacy instrument used.

Determining Item Fit

An item is considered to fit if the p-value of the χ^2 is greater than 0.05 (Paek & Cole, 2019). The results of the items that fit the Rasch model can be seen in Table 6. Item 3 did not fit, and therefore, it was not used in subsequent tests and predicting participants' scores.

Table 6. Item Fit

Item	Difficulty (b)	χ^2	P_value	Item Status
Butir 1	-0.61	7.47	0.19	Fit
Butir 2	-1.56	2.65	0.62	Fit
Butir 3	-1.15	14.79	0.02	Tidak Fit
Butir 4	0.45	5.07	0.53	Fit
Butir 5	-0.96	2.23	0.90	Fit
Butir 7	-0.61	4.85	0.43	Fit
Butir 8	-0.79	1.11	0.95	Fit
Butir 9	-0.61	6.05	0.30	Fit
Butir 11	0.54	8.04	0.23	Fit
Butir 12	-0.53	5.28	0.38	Fit
Butir 13	-0.20	12.54	0.05	Fit
Butir 14	-0.87	6.28	0.28	Fit
Butir 15	-1.25	6.08	0.30	Fit

Function Of Information Of The Mathematics Literacy

The function of information of the mathematics literacy instrument indicates the extent to which the test can be effectively used for a certain ability. The shaded area in Figure 5 shows the intersection of the information function and the standard error.

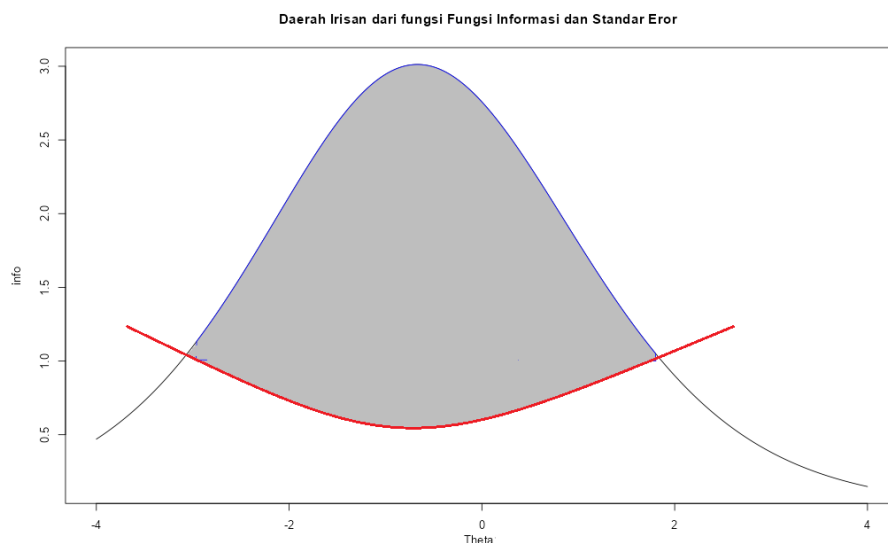


Figure 5. Information Function Curve and Standard Error test.

In Figure 5, it can be seen that the intersection coordinates of the information function and the standard error are $(-2.965, 1.007)$ and $(1.805, 1.001)$. The abscissas of these intersection coordinates explain the suitable ability range for taking this test or the suitability of this test instrument for use on students with abilities from -2.965 (low) to 1.805 (moderate). Furthermore, the results of the R package *mirt* calculation show that the total information value and standard error are 12 and 0.29, respectively.

DISCUSSION

The unidimensionality result explains that the mathematics literacy instrument used measures one dimension or one trait. The trait referred to in this article is the mathematics literacy ability of students in algebra content. The instrument's construct in algebra content that is measured has a construct that is suitable for the CFA order-2 model, where mathematics literacy ability can be measured by its constituent sub-elements.

Local independence is one of the assumptions in IRT that statistically requires that each item does not have a strong relationship with each other. Violations of this assumption will affect the estimation of ability parameters (Edwards et al., 2018). In Table 3, the absolute value of Q3 statistically indicates the occurrence of local dependence or that the local independence assumption is not met. According to (Toland, 2014), the item that causes local dependence should be removed, and IRT analysis should be performed from the beginning of the procedure. According to (Edwards et al., 2018) and (Edelen & Reeve, 2007), local dependence can occur on items that have similarity in the content of the test used and have the same solution procedure. Furthermore, (Edwards et al., 2018) gives an example of two items, x_1 and x_2 , that cause local dependence. When a test participant responds incorrectly to x_1 and correctly to x_2 , the estimation of the participant's ability will be obtained from item x_2 , and vice versa. Based on the opinion of (Toland, 2014), in the example given by (Edwards et al., 2018), one item must be removed (not used) because it can cause bias in the estimation of ability. However, if x_1 and x_2 theoretically do not have a relationship in concept, content, and procedure, then one item does not have to be removed. Because if one item is responded incorrectly, it will still provide different information on the estimation of ability. This is the reason why the author disregarded the results of the local independence assumption testing in Table 4.

The mathematical literacy instrument for algebra content has been shown to satisfy the parameter invariance property, even though the sample size used was small. The fulfillment of this assumption indicates that the item parameters are not influenced by student ability, and conversely, student ability is not influenced by the test item parameters. Thus, the instrument can be used to measure mathematical literacy in algebra content in other samples in the population. The assumption of item invariance theoretically applies in all cases, but in real life, data does not always support it. This may be because poorly written items are interpreted differently by different samples (Nguyen et al., 2014).

Table 6 shows that every item fits the Rasch model used, except for item 3. The Rasch model only produces difficulty parameters, which can be seen in column 2 of Table 6. Wulandari et al. (2020) categorized the level of difficulty into three categories, namely difficult ($b \geq 0.7$), moderate ($-1 \leq b < 0.7$), and easy ($b < -1$). The mathematical literacy instrument for algebra content in Table 6 has a level of difficulty at the easy level for items 2 and 15. Apart from these two items, they have a moderate level of difficulty. The difficulty level parameters for each item in Table 6, except for item 3, will be used for scoring or estimating students' mathematical literacy abilities.

The information function generated from calibration provides information about the test. This information function is formed from the information functions of each item except for item 3. Figure 6 explains that the test is suitable for measuring mathematical literacy abilities in students who have abilities between -2.965 (low) to 1.805 (moderate). These results indicate that the algebra content mathematical literacy instrument used can be used to measure students who are not in the sample.

These calibration results are then used for scoring students' mathematical literacy abilities. Score estimation is done using item parameters (Brown & Croudace, 2014) obtained in Table 6. Scoring is based on student responses to each mathematical literacy test item. According to Brown (2018), IRT will estimate scores for each student based on their responses to each test item. In this article, the estimation method used is the EAP method. The results of scoring for students who are part of the sample using the 12 items that fit in Table 6 can be seen in Figure 7.

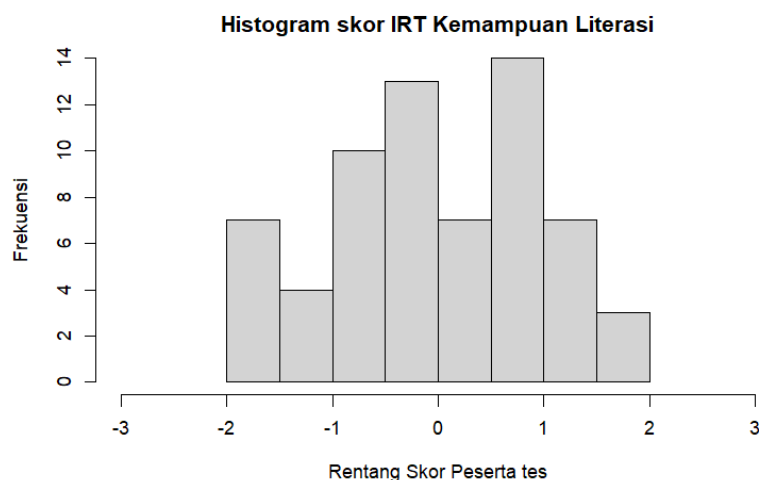


Figure 7. Histogram of Students' Mathematical Literacy Ability

As described in the method section, this study used a test consisting of 15 items. After validation and calibration, only 12 items were found to be usable. Each item was in the form of multiple choice with 4 options. Considering the dichotomous nature of the responses (correct/incorrect), there were $2^{12} = 4096$ possible response patterns. However, the data used in this study only came from 65 students. Assuming that these 65 students produced different response patterns, there were still 4031 response patterns that were not used in estimating the item parameters using the Rasch model. Interestingly, based on the item parameters obtained above, IRT was still able to estimate the scores of students who had response patterns that were not used in parameter estimation (Brown & Croudace, 2014). Thus, this algebraic literacy instrument can be used to estimate the scores of students who were not included in the

sample. As shown in Table 7, score estimation was performed for ten students whose response patterns were different from those in the sample.

Table 7. Simulation of Mathematics Literacy Score Estimation Using IRT Rasch Model

Name	item													Score IRT	Score 0-100
	1	2	3	4	5	6	7	8	9	10	11	12	13		
Student 1	1	0	1	0	0	0	0	0	0	0	0	0	0	-1.561	23.98
Student 2	1	1	0	0	0	0	0	0	0	0	0	0	0	-1.561	23.98
Student 3	1	0	0	0	0	0	0	1	0	0	1	0	0	-1.26	29.00
Student 4	1	1	0	0	1	0	0	1	0	0	0	1	1	-0.697	38.38
Student 5	1	1	1	1	1	1	1	0	0	0	0	0	0	-0.143	47.62
Student 6	1	1	1	1	0	0	0	0	0	0	0	1	1	-0.697	38.38
Student 7	1	1	1	1	1	1	1	1	0	0	0	0	0	0.146	52.43
Student 8	1	1	1	1	1	1	1	1	1	0	0	0	0	0.453	57.55
Student 9	1	1	1	1	1	1	1	1	1	1	0	0	0	0.786	63.10
Student 10	0	0	0	0	0	0	0	0	0	0	0	0	0	-2.247	12.55

The weaknesses of this study are (1) the limited sample size used. Therefore, there is no validation data used for the application of estimating literacy ability scores using the item parameters obtained. The small sample size also affects the method of proving the invariance used in this article, namely the split-sample method. (2) The method of proving parameter invariance only uses one possibility, namely the odd-even split-sample. Based on these weaknesses, it is hoped that in further research, a larger sample size can be used and other methods can be used to prove the invariance property.

CONCLUSION

Based on the results and discussions, the algebra content mathematics literacy instrument used has characteristics that fit the Rasch model. As a result, the item characteristics inherent in the instrument are only in the form of difficulty parameters. Two items are in the easy category, 10 items are in the moderate category, and the rest cannot be used. The intersection of the information function and standard error provides information that this literacy instrument will provide very accurate information if used on samples or students with literacy abilities between -2.965 to 1.085. Furthermore, the item parameters can be used to score both the used sample and theoretically on other samples in the population.

BIBLIOGRAPHY

- Brown, A. (2018). Item Response Theory Approaches to Test Scoring and Evaluating the Score Accuracy. In *The Wiley Handbook of Psychometric Testing* (pp. 607–638). Wiley. <https://doi.org/10.1002/9781118489772.ch20>
- Brown, A., & Croudace, T. (2014). Scoring and Estimating Score Precision Using Multidimensional IRT Models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 325–351). Routledge. <https://doi.org/10.4324/9781315736013-26>
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>
- Dewanti, S. S., Hadi, S., & Nu'man, M. (2021). The Application of Item Response Theory in Analysis of Characteristics of Mathematical Literacy Test Items. *İlköğretim Online*, 20(1). <https://doi.org/10.17051/ilkonline.2021.01.119>
- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire

- development, evaluation, and refinement. *Quality of Life Research*, 16(S1), 5–18. <https://doi.org/10.1007/s11136-007-9198-0>
- Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23(1), 138–149. <https://doi.org/10.1037/met000121>
- Finch, H. (2014). Measurement Invariance. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 3909–3912). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_1759
- Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). Multivariate Data Analysis, Multivariate Data Analysis. In *Multivariate Data Analysis*.
- Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *fundamental of item response theory*. SAGE.
- Hasanah, M., & Hakim, D. L. (2022). Kemampuan Literasi Matematis Pada Soal Matematika PISA Konten Quantity dan Konten Change and Relationship. *JURING (Journal for Research in Mathematics Learning)*, 5(2), 157. <https://doi.org/10.24014/juring.v5i2.13785>
- Hertiandito, L. T. (2016). Kemampuan literasi matematika siswa SMP pada pembelajaran Knisley dengan tinjauan gaya belajar. *PRISMA, Prosiding Seminar Nasional Matematika, 2011*.
- Kemendikbud. (2019). Kajian Akademik dan Rekomendasi Reformasi Sistem Asesmen Nasional. *Kementerian Pendidikan Dan Kebudayaan*.
- Kemendikbud. (2021). Asesmen Nasional: Lembar Tanya Jawab. *Kementerian Pendidikan Dan Kebudayaan*.
- Larasaty, B. M., Mustiani, & Pratini, H. S. (2018). Peningkatan Kemampuan Literasi Matematika Siswa Kelas VIII SMP Bopkri 3 Yogyakarta Melalui Pendekatan PMRI Berbasis PISA Pada Materi Pokok SPLDV. *Prosiding Seminar Nasional Etnomatnesia*.
- Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software*, 76(Code Snippet 1). <https://doi.org/10.18637/jss.v076.c01>
- Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-69218-0>
- Mansur, N. (2018). Melatih Literasi Matematika Siswa dengan Soal PISA. *Prisma*, 1.
- Masjaya, & Wardono. (2018). Pentingnya Kemampuan Literasi Matematika untuk Menumbuhkan Kemampuan Koneksi Matematika dalam Meningkatkan SDM. *PRISMA, Prosiding Seminar Nasional Matematika, 1*.
- Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective*, 11(3), 71–101. <https://doi.org/10.1080/15366367.2013.831680>
- Maydeu-Olivares, A. (2014). Evaluating the Fit of IRT Models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 129–145). Routledge. <https://doi.org/10.4324/9781315736013-15>
- Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An Introduction to Item Response Theory for Patient-Reported Outcome Measurement. *The Patient - Patient-Centered Outcomes Research*, 7(1), 23–35. <https://doi.org/10.1007/s40271-013-0041-0>

- OECD. (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*.
- Paek, I., & Cole, K. (2019). *Using R for Item Response Theory Model Applications*. Routledge. <https://doi.org/10.4324/9781351008167>
- R Core Team. (2022). *R: A Language and environment for statistical computing*. R Foundation for statistical Computing. <https://www.R-project.org/>.
- Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya*.
- Retnawati, H. (2016). *validitas dan reliabilitas dan karakteristik butir* (1st ed.). Parama Publising.
- Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). COMPARISON OF MULTISTAGE TESTS WITH COMPUTERIZED ADAPTIVE AND PAPER-AND-PENCIL TESTS. *ETS Research Report Series*, 2007(1), i–27. <https://doi.org/10.1002/j.2333-8504.2007.tb02046.x>
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement*, 66(1), 63–84. <https://doi.org/10.1177/0013164404273942>
- Susanto, H. P. (2022). *package hpsCFA untuk Analisis Validitas Kontruk menggunakan CFA*. <https://github.com/SusantoHP/hpsCFA>
- Tabachnick, L. S., & Fidell, B. G. (2014). *Using multivariate statistics: Pearson New International Edition*.
- Toland, M. D. (2014). Practical Guide to Conducting an Item Response Theory Analysis. *The Journal of Early Adolescence*, 34(1), 120–151. <https://doi.org/10.1177/0272431613511332>
- Vale, P., Murray, S., & Brown, B. (2013). Mathematical literacy examination items and student errors: An analysis of English Second Language students' responses. *Per Linguam*, 28(2). <https://doi.org/10.5785/28-2-531>
- Wulandari, F., Hadi, S., & Haryanto, H. (2020). Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability. *Computer Science and Information Technology*, 8(3), 66–73. <https://doi.org/10.13189/csit.2020.080302>
- Yuberta, K. R., Nari, N., & Gustia, E. (2020). KEMAMPUAN LITERASI MATEMATIS SISWA DENGAN MENERAPKAN MODEL PEMBELAJARAN CREATIVE PROBLEM SOLVING (CPS). *Jurnal Sainika Unpam: Jurnal Sains Dan Matematika Unpam*, 3(1), 68. <https://doi.org/10.32493/jsmu.v3i1.6269>