Ilkogretim Online - Elementary Education Online, 2021; 20 (1): pp. 1226-1237
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.119

The Application of Item Response Theory in Analysis of Characteristics of Mathematical Literacy Test Items

Sintha Sih Dewanti, *Universitas Negeri Yogyakarta & UIN Sunan Kalijaga Yogyakarta, Indonesia, sinthasih.2018@student.uny.ac.id & sintha.dewanti@uin-suka.ac.id ORCHID: 0000-0001-5966-1354*
Samsul Hadi, *Universitas Negeri Yogyakarta, Indonesia, samsul_hd@uny.ac.id ORCHID: 0000-0003-3437-2542*
Mulin Nu'man, *Universitas Negeri Yogyakarta & UIN Sunan Kalijaga Yogyakarta, Indonesia, mulinnuman.2019@student.uny.ac.id & mulin.nu@uin-suka.ac.id ORCHID: 0000-0002-7046-9408*
Ibrahim, *UIN Sunan Kalijaga Yogyakarta, Indonesia, ibrahim@uin-suka.ac.id ORCHID: 0000-0001-6945-0040*

Abstract. This study aims to analyze the mathematical literacy abilities of students using IRT with the GPCM approach. The results of the model compatibility test show that the GPCM is suitable for analyzing mathematical literacy test instruments with 4 levels of mathematical literacy. The results of the item parameter estimation show that there is 1 level of mathematical literacy which has a high category for item discrimination, while the other 3 levels have a medium category. Based on item difficulty levels, 1 level of mathematical literacy has a difficult category, while 3 levels have easy categories. Most students are at the level of ability $0 < \theta < 1$, and there are 52% of students above the average mathematical literacy. Based on the test information function and the standard error shows that the mathematical literacy instrument is suitable for students with the ability $-2 < \theta < -0.5$.

Keywords: Mathematical literacy, Characteristics of items, IRT, GPCM

Received: 03.11.2020	Accepted: 02.12.2020	Published: 05.01.2021

INTRODUCTION

Mathematics as a compulsory subject is expected to provide students with the ability to reason and analyze in solving daily problems (Gravemeijer, Stephan, Julie, Lin, & Ohtani, 2017). Paulos (Nickerson, 2011) argues that mathematics is not a matter of entering numbers into formulas and performing memorization procedures. However, mathematics is a way of thinking and questioning something foreign to students. Van de Walle, Karp, & Bay-Williams (2010) revealed that doing mathematical activities means producing strategies to solve problems, applying an approach, investigating the process of whether it leads to a solution, and checking whether the resulting answer makes sense. Mathematical activities in class must be done carefully in modeling real-world things into mathematics (Blum & Ferri, 2009). Related to that, Schoenfeld (1992) holds that mathematics is used as a tool to understand the patterns that exist in the world around us, as well as the patterns that exist in our minds. In this opinion, mathematics can be interpreted as a broad science of pattern searching (Resnik, 1997).

In line with this opinion, Kilpatrick & Swafford (2002) argues that mathematical skills involve five interrelated abilities, namely: 1) understanding mathematical concepts, 2) calculating fluently, 3) applying mathematical concepts to solve problems, 4) reasoning logically, and 5) involved with mathematics, seeing mathematics as something that makes sense, is useful, and can be done. These five abilities will appear fully in the process of solving real problems. This is in line with the view of NCTM (Reys, Lindquist, Lambdin, & Smith, 2009), problem-solving, reasoning and proof, communication, and representation are used as standard processes in learning mathematics. Such mathematical abilities are known as mathematical literacy abilities.

In PISA (OECD, 2013), mathematical literacy refers to the ability of students to formulate a problem, use mathematical concepts, and interpret mathematical problem-solving in various contexts. This activity involves the process of mathematical reasoning using mathematical concepts, procedures, facts, or tools to provide an overview, explanation, and prediction of a phenomenon. Students who have good mathematical literacy skills will better understand the role of mathematics so that it will provide strong support in making appropriate assessments and decisions (Jablonka & Niss, 2014). This shows that mathematical literacy does not only play a role in the process of mastering mathematical material, but mathematical literacy is also used in the process of using mathematical reasoning, concepts, and tools in solving everyday problems (Österman & Bråting, 2019).

Based on this description, it is clear that mathematical literacy is an important ability in learning mathematics. So far, the estimation of mathematical literacy ability is based on the results of an analysis of the responses or answers given by students globally by using the Classical Test Theory (CTT). CTT has been widely used in the field of measurement until now. However, CTT has various limitations including: 1) the characteristics of test items in the form of difficulty and discrimination of items depending on the characteristics of test-takers; 2) the ability of the test takers to depend on the characteristics of the items being tested; 3) error score estimation applies to all test takers; 4) can not provide information about the response to each item; and 5) using parallel assumptions that are sometimes difficult to fulfill, such as mean and variance must be the same (Fan, 1998; Lawson, 2006; Mardapi, 2008; Stage, 2003). The same thing was stated by Hambleton and Swaminathan (1985) that the assumption of parallel tests on CTT is difficult to fulfill, and does not provide information about the ability of test-takers.

In addition to CTT, there is an Item Response Theory (IRT) that can be used to analyze mathematical literacy abilities. The difference between the two analyzes is the focus of the information provided (Janssen, Meier, & Trace, 2014). CTT focuses on test level information, while IRT focuses on item level information (Bichi, Embong, Mamat, & Maiwada, 2015). Therefore, it is expected that IRT can cover the deficiencies found in CTT. The most important thing in IRT is the determination of response models or item characteristics. The response model must meet several assumptions underlying IRT, namely: 1) local independence, meaning that the opportunity to answer one item correctly is not influenced by the opportunity to answer another item correctly; 2) unidimensional, meaning that the test measures one dimension of ability; and 3) parameter invariance, meaning that the response pattern of each test item can be described in the form of item characteristic curves (Hambleton, Swaminathan, & Rogers, 1991; Naga, 1992). IRT built a model that connected the characteristics of items with the characteristics of the participants. With several specific conditions, this relationship model is made to apply freely to any item groups and groups of participants who meet these requirements. This study aims to analyze empirically the characteristics of mathematical literacy ability tests based on IRT.

METHODS

This research is a quantitative descriptive study, which aims to analyze empirically the characteristics of mathematical literacy ability tests based on IRT. The data collection of mathematical literacy skills was carried out through a written test with an allocation of 60 minutes to 258 students. The test in the form of description consists of 4 items of mathematical literacy questions, each of which represents 4 levels of mathematical literacy from level_1 to level_4. The test is arranged with reference to the level and indicators of mathematical literacy abilities, which are associated with mathematics material class VII, namely lines and angles, mathematical comparison, Cartesian coordinates, and scaling.

The polytomous item response model is used to scale students' responses to a test item. The variety of test data generated by polytomous items can be grouped into two, namely: 1) tests with all items having many of the same response categories, and 2) tests with items having many diverse response categories (Nandakumar, Yu, & Zhang, 2011). In this study, scoring polytomous items using 7 scales (0 to 6) was carried out by looking at the stages of the test participants at each level of mathematical literacy. Some researchers say the use of 7 scales will maximize the reliability of internal consistency, but some mention 4 scales and another 3 scales (Chang, 1994). Linn and Gronlund (Boughton, Klinger & Gierl, 2001) recommend the use of a scale of 3 – 7 categories.

The mathematical literacy test allows us to have many different response categories for each item. This is due to the operational level of mathematical literacy ability on each item is different. There are several polytomous models on the IRT (Van der Linden & Hambleton, 1997) including the Rating Scale Model (RSM), Graded Response Model (GRM), Partial Credit Model (PCM), and Generalized Partial Credit Model (GPCM). Related to discriminant parameters, the Rasch model assumption (PCM) where inter-items have the same discrimination value on empirical data is usually violated (Ware, Bjorner & Kosinski, 2000). RSM cannot be used if the item score category varies among items and PCM is a model with a fixed discriminant factor value for all items. Therefore, the item response model that might be used is GRM or GPCM. In this study, data analysis techniques were performed using the IRT model with the GPCM approach and computation using the R software.

In GPCM, the probability of a test participant having a k category is explained that someone with a certain level of ability reaches the k score category beyond the (k – 1) score category (Tang, 1996). If there is an item that is polytomous scored having m score categories, then the higher score category represents a higher ability than (m – 1). However, the level of difficulty in the second stage does not have to be higher than in the first stage, and vice versa (De Ayala, 1993). According to Muraki (1999), GPCM is a common form of PCM. GPCM is similar to PCM in that it conceptualizes the choice of characteristic curve options.

GPCM is also similar to the 2PL model where item discrimination parameters can vary in each item. GPCM estimates the unique item discrimination parameters for each item. The item discrimination parameter indicates the extent to which category responses vary between items when the latent trait changes (Muraki, 1992).

Opportunities to obtain the X_{ik} category (X_{ik} = 0, 1, 2, …, m_i) in item i for GPCM (De Ayala, 2013) can be written as follows.

$$P(X_i \mid \theta, a_i, \delta_{ik}) = \frac{\exp \sum_{h=0}^{X_{ik}} a_i(\theta - \delta_{ih})}{\sum_{c=1}^{m_i} \exp \sum_{h=1}^{c} a_i(\theta - \delta_{ih})}$$

In the formula $P(X_i \mid \theta, a_i, \delta_{ik})$ it is explained that θ is a latent trait, $a_i$is a discrimination parameter for item i, δ_{ik} is a step parameter (difficulty step) that represents the relative difficulty in obtaining the category k beyond the category (k – 1). The threshold (δ_{ik}) in GPCM is similar to PCM, that is, the threshold is not limited in the same order as the response category.

RESULTS AND DISCUSSION

In conducting data analysis using item response theory, the first thing to do is to examine the dimensions of the empirical data obtained. It is important to ensure unidimensionality, where only one latent attribute can explain the whole matrix of test participants' responses (Lord, 1980). Information about the dimensionality of this test will also provide structural evidence related to the consistency between the internal structure of the test and the construct structure (Fiske, 2002). Then information about the structure of these dimensions can be used as a foundation in reporting scores or subscales. Multidimensionality will occur when tests are designed to measure complex latent attributes (Camilli, Wang, & Fesq, 1995). If a test is designed to measure complex latent attributes, it is difficult to claim that the construct measured is pure unidimensional.

Conditioning so that scores are comparable between groups or between times should be a serious concern because it involves validity, especially aspects of generalization (Messick, 1995). Structural differences between groups or time can be traced based on their dimensionality (Tate, 2002). Formally, the dimensionality of a test can be defined as the minimum number of dimensions that can explain to data and models so that they are Monotone Locally Independent (MLI) (Stout, 2002). The testing process is carried out by the Exploratory Factor Analysis (EFA) with the principal component method. The initial assumptions that must be met in the EFA are the KMO and Bartlett tests. The KMO and Bartlett Test outputs are presented in Table 1.

Table 1. *KMO and Bartlett test outputs*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.783
Bartlett's Test of Sphericity	Approx. Chi-Square	315.157
	df	6
	Sig.	0.000

The results of the KMO Test are KMO > 0.5, which indicates that the sample size of 258 already has sufficient data. Also, the significance of Bartlett's test (sig. < 0.05) indicates that the H_0 (correlation matrix is the identity matrix) is rejected so that the data form a correlation matrix with a close relationship between variables. Then based on anti-images correlation, Measures of Adequate Sampling (MSA) > 0.5 so that all data are feasible for factor analysis. The Anti-image Matrices output is presented in Table 2.

Table 2. *Anti-image matrices*

Anti-image Correlation		**Item_1**	**Item_2**	**Item_3**	**Item_4**
	Item_1	0.767[a]	-0.321	-0.240	-0.322
	Item_2	-0.321	0.769[a]	-0.327	-0.225
	Item_3	-0.240	-0.327	0.802[a]	-0.063
	Item_4	-0.322	-0.225	-0.063	0.807[a]
a. Measures of Sampling Adequacy (MSA)					

Principal Component Analysis (PCA) can be used to minimize the number of variables observed so that a small number of main components is formed from most of the variance of the observed variables. The number of factors can be determined by selecting factors that have an eigenvalue greater than 1. Eigenvalues from the PCA are presented in Table 3.

Table 3. *Eigenvalues of principal component analysis*

Component		Initial Eigenvalues		
		Total	**% of Variance**	**Cumulative %**
	1	2.512	62.793	62.793
	2	0.633	15.830	78.622
	3	0.443	11.080	89.702
	4	0.412	10.298	100.000

The results of the factor analysis show that there is 1 factor that has an eigenvalue of more than 1. It can be said that the 4 items analyzed are grouped into 1 factor which can explain about 62.79% of the total variance, which means that these factors are more than the average total variance in the items (Guttman, 1954; Kaiser, 1960). The first-factor eigenvalue is several times the second-factor eigenvalue, then the second-factor eigenvalue and so on are almost the same. Therefore, mathematics literacy tests meet the requirements of unidimensional (Naga, 1992). The unidimensional nature can be described more clearly by the scree plot by finding the elbows (bends) in Figure 1.

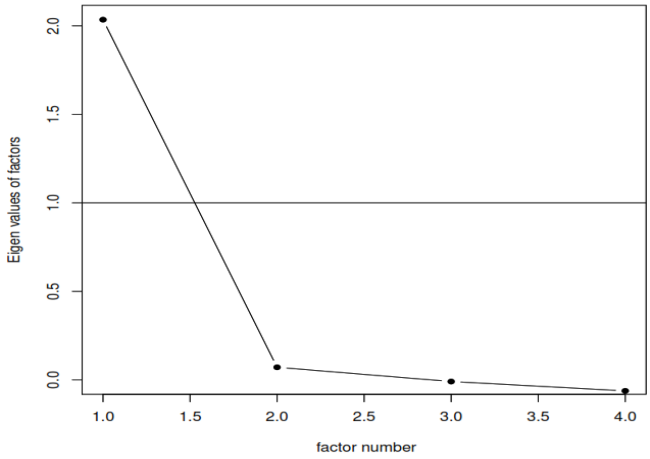

Figure 1. *Scree plot of principal component analysis*

Thus, empirical data shows that items of mathematical literacy only measure one factor (unidimensional). Dimensionality in measurement can also be interpreted as the number of latent attributes that underlie the ability of test-takers to respond to test items (Chou & Wang, 2010). In the context of ability tests, dimensionality is referred to as the number of abilities measured by a test or by a collection of items. When related to the content of the test material, dimensionality can be seen as aspects of measurement designed to be measured by the test (Mislevy, Almond, & Luke, 2003) or can also be seen as an analysis of the response data on a set of items (Reckase, 2009). This dimensional analysis is an EFA that is used to identify the relationship between manifest variables or indicator variables in constructing mathematical literacy constructs.

The results of EFA with the principal axis factoring and rotation = varimax methods obtain a factor loading of each item on one factor or dimension. EFA results are shown in Figure 2. Factor loading is a coefficient that explains the level of relationship of items with latent variables in the form of mathematical literacy. The EFA results obtained factor loading at intervals of 0.63 – 0.78 (factor loading > 0.30), which means that instrument items can correctly interpret mathematical literacy (Brown, 2015; Harrington, 2009; Thompson, 2004).

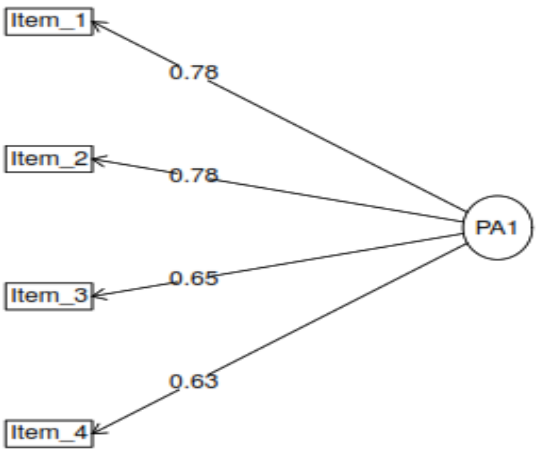

Figure 2. *Factor loadings from exploratory factor analysis results*

The development of test instruments must also pay attention to test reliability. Crocker and Algina (1986) describe reliability as a measure of instrument consistency in the resulting score. Reliability means the extent to which the results of measurement have credibility, reliability, constancy, consistency, stability that can be trusted (Chakrabartty, 2013). The consistency of the test in question is if the measurement of the same attribute is repeated then the test will give identical or very similar condition results. Reliability refers to measurements whose results are consistent with the same value (Blumberg, Cooper, & Schindler, 2014). This shows the extent of measurement without bias by ensuring consistent measurement times. A set of tests is said to be reliable if it has a high correlation between the observed score and the actual score (Allen & Yen, 1979). Reliability (ρ) of a test is generally expressed numerically in the form of coefficients of magnitude $-1 \leq \rho \leq 1$. The higher the reliability coefficient, the more accurate the results will be, which means increasing the opportunity to make correct decisions in this study. The reliability test on this mathematical literacy test instrument uses Cronbach's alpha technique and produces a reliability coefficient of 0.796 which is included in the high category.

Furthermore, the data were analyzed with unidimensional item response theory for polytomous items, using GPCM. The S_X^2 test was chosen to test item-fit in this study because the S_X^2 item-fit index is suitable for polytomous IRT model items in educational and psychological testing programs (Kang & Chen, 2007). The S_X^2 index can also be generalized and applied to a good-of-fit test for polytomous items, such as GPCM (Roberts, 2008). Model match test results are presented in Table 4. Based on the value of p. $S\_X^2 > 0.05$ for each item shows that GPCM is suitable for analyzing mathematical literacy abilities.

Table 4. *Results from the model compatibility test*

Item	S_X^2	df.S_X^2	RMSEA.S_X^2	p.S_X^2
Item_1	25.439	22	0.025	0.277
Item_2	34.493	26	0.036	0.123
Item_3	34.287	28	0.03	0.192
Item_4	25.17	24	0.014	0.397

Item parameter estimation using GPCM will reveal two-item parameters, namely difficulty level and distinguishing power. The parameter a explains how much the item can distinguish between individuals with different abilities. The parameter b is interpreted as the relative difficulty of a step compared to other steps in an item. The parameter b_i can be interpreted as the estimated value of the difficulty level parameter to reach the value category i. Scoring each item uses 7 scales, so each item has 6 item difficulties (b_i, with i = 1, 2, 3, ..., 6). The results from the estimated parameters of mathematical literacy items are presented in Table 5.

Table 5. *Item parameter estimation using GPCM*

Item	a	b	b_1	b_2	b_3	b_4	b_5	b_6
Item_1	1.543	-0.694	-2.793	-1.812	-1.004	-0.972	0.968	1.447
Item_2	1.173	-0.821	-2.276	-2.244	-0.909	-0.381	0.406	0.480

Item_3	0.804	-0.765	-1.902	-1.906	-0.767	-1.792	-0.115	1.892
Item_4	0.896	0.803	-0.885	0.508	0.736	1.588	1.143	1.728

There are 3 steps in 2 items that have an item difficulty level below -2.0, but for an average of item difficulty levels are good. Similarly, the distinguishing power of each item is quite good. An item is classified as good if it has a discrimination index of items 0.0 to 2.0 and an index of difficulty -2.0 to 2.0 (Hambleton & Swaminathan, 1985). Generally, applies $b_i < b_{i+1}$, because students will complete a step if they have completed the previous step. Item_3 and the scoring need to be reviewed, because item_3 applies $b_2 < b_1$ and $b_4 < b_3$.

Based on the item discrimination index, item_1 has the highest discrimination index compared to other items. This means that item_1 can distinguish students' mathematical literacy abilities well. Item_1 indicator is the student can determine the angle rotation in degrees in the Ferris wheel problem. The level of mathematical literacy ability on item_1 is level_1. At level_1, students are expected to be able to answer questions in a general context that is familiar to students. All information relevant to solving the questions is available on the question points. Based on explicit question instruction, students can identify information, then use routine procedures to solve it. Items with explicit instruction are effectively used to help students who have difficulty learning mathematics. According to Doabler and Fien (2013), explicit instruction can be used by teachers to facilitate students in understanding critical mathematics content. This can help students take action by the stimulus provided.

Item_3 has the lowest discrimination index compared to other items but is still included in the medium category. Item_3 deals with Cartesian coordinates with problems determining the position of objects. The level of mathematical literacy ability on item_3 is level_3. At level_3 students are expected to be able to carry out procedures well, including procedures that require decisions in sequence. It is important to understand how the order of decisions made by students that will affect observations and optimal results based on criteria (Maillard, 2019). Students can select and implement a variety of problem-solving strategies ranging from simple strategies to complex strategies (Gick, 1986). Students can interpret and use representations based on different sources of information and can then state the reasons. Also, students can communicate the results of interpretations and their reasons. Giving a reason can be used to investigate students' construction errors in mathematical representation. Construction errors experienced by students can occur in various forms, namely loss of representation attributes, mapping from one representation attribute to another, mixing of two different schemes that appear simultaneously, disconnecting connections at the initial coordination stage, and implementation errors (Afriyani, Sa'dijah, Subanji, & Muksar, 2019).

Based on the level of difficulty, item_2 has the lowest value compared to other items, meaning that item_2 is easiest. The problem with item_2 is related to the comparison material, i.e. determine the time needed to spend the medicine according to the rules. The level of mathematical literacy ability on item_2 is level_2. At level_2, students are expected to be able to interpret information and recognize situations in a mathematical context that require direct inference. Students sort the relevant information from a single source and use a single representation. Students can work on basic algorithms, use formulas, carry out procedures, or simple conversions. Students can provide reasons directly and make interpretations. Some of the interpretations that students can do include: 1) interpreting simple text and linking it correctly to graphic elements; 2) interpret a simple text containing a simple algorithm and then apply it; 3) interpret a simple text by using proportional reasoning or doing calculations; and 4) interpret simple patterns (OECD, 2004).

Item_4 has the highest item difficulty level compared to other items. Item_4 deals with scaling material, i.e. calculating the area of an area based on a map using the concept of scale. The level of mathematical literacy ability in item_4 is level_4. At level_4, students are expected to be able to do mathematical modeling effectively in complex concrete situations. Students can select and integrate different representations, and relate to real situations. Students can use their skills well in suggesting a reason and a flexible viewpoint according to the context. Students can provide explanations and communicate them with arguments based on their interpretations and actions.

Categorical Response Function (CRF) curves indicate the likelihood of respondents choosing a particular score on a scale (0 – 6) at various levels of latent traits. An item will be better at distinguishing the abilities of each individual if the curve peaks and spreads at all levels of latent traits. The relationship of the probability of answering correctly for each ability is presented in the Categorical Response Function (CRF) curve in Figure 3.

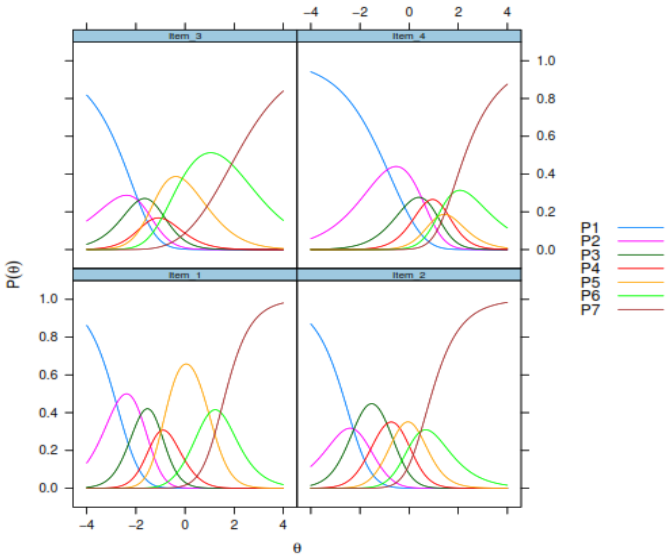

Figure 3. *Categorical response function graph*

GPCM is formulated based on the assumption that each probability of choosing the k-th category beyond the (k – 1) category is built by a dichotomous model (Muraki, 1992; 1993). The parameter b_{jk} is the intersection point between the $P_{jk}(\theta)$ curve and $P_{jk-1}(\theta)$ curve. The two curves only intersect at one point on the scale of θ (Van der Linden & Hambleton, 1997) with 3 possibilities as follows: 1) If $\theta = b_{jk}$, then $P_{jk}(\theta) = P_{jk-1}(\theta)$; 2) If $\theta > b_{jk}$, then $P_{jk}(\theta) > P_{jk-1}(\theta)$; and 3) If $\theta < b_{jk}$, then $P_{jk}(\theta) < P_{jk-1}(\theta)$, k = 1, 2, 3, ..., m_j. P_{jk} is a special probability of choosing the k category from the m_j +1 category.

The threshold shows the meeting point of two category probability lines in one item. The threshold is a point where two categories have the same probability to be chosen by the related level of the trait (Linacre, 2006). The individual's probability of responding to category x at this stage m_i is the difference between the trait level (θ) and the threshold (δ_{ij}). The intersection of P_1 and P_2 shows the minimum ability students must have to get a score of 1, the intersection of P_2 and P_3 shows the minimum ability a student must have to get a score of 2, and so on. In general, it can be written that the intersection of P_i and P_{i+1} shows the minimum ability students must have to get a score of i. A good item if the intersection of P_i and P_{i+1} is to the left of the intersection of P_{i+1} and P_{i+2}, which means the ability that must be possessed to get i score is lower than the ability that must be had to get a score of (i +1) (Embretson & Reise, 2000). This applies to item_1, item_2, and item_4, but does not apply to item_3. However, the value of δ_{ij} does not always have to be sequential on item i because δ_{ij} is a relative magnitude of two adjacent probabilities (De Ayala, 1993; Muraki, 1992). The threshold can also be interpreted as a point on the latent trait scale, where the response curves will intersect for two consecutive categories.

There are several approaches to estimating mathematical literacy abilities, including Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP) (Embretson & Reise, 2000). MLE is a general method for estimating model parameters, is quite effective with large samples and valid model applications (Longford, 2008). MLE has many optimal properties in the estimation, namely: information sufficiency, data consistency, efficiency, and invariant parameterization (Myung, 2003). The highest odds will depend on the probability of correct answers and wrong answers by participants, and also on the logistical parameters used. Thus, determining the value of maximum capability is done through iteration calculations (Baker, 2001). MAP estimation is almost the same as MLE, only MAP calculation using Fisher scoring using posterior information (Han, 2016). The EAP method is based on the Bayes theorem which combines the previous distribution with the sample distribution (Baker, 1991; Bock & Aitkin, 1981).

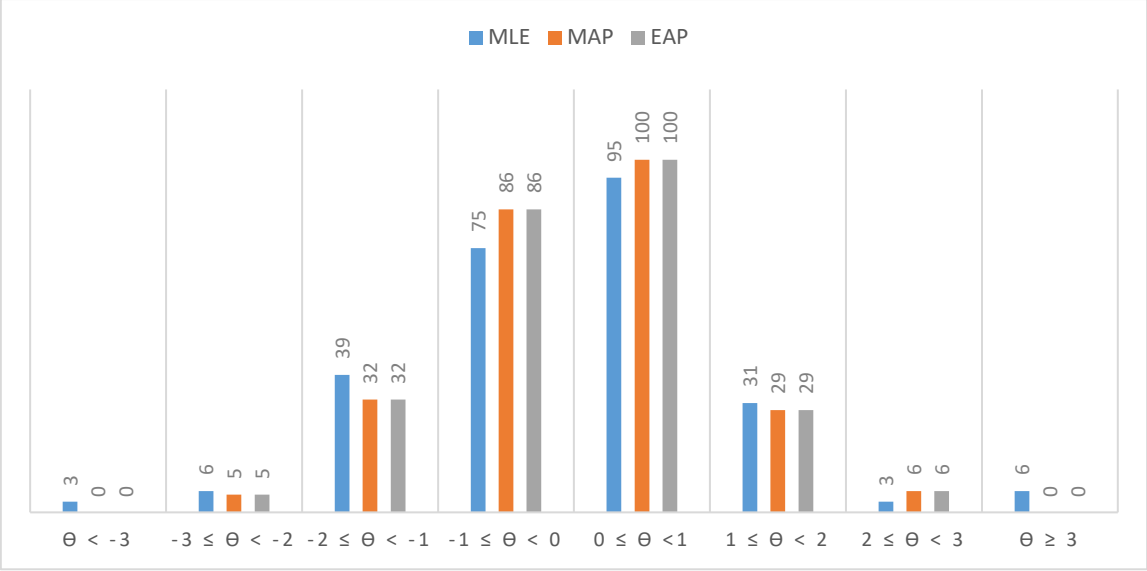

Figure 4. *Histogram of the mathematical literacy ability distribution*

The results of the test taker's ability estimation with all three approaches are presented in Figure 4. The mathematical literacy abilities of students spread normally. This is consistent with the theory that the more respondents, the estimated distribution of ability will approach the normal distribution. MAP and EAP produce estimates of the same mathematical literacy ability, while MLE produces slightly different estimates of ability. Estimated literacy abilities of students with the MAP and EAP approaches are at intervals of $-3 < \theta < 3$, while the estimated abilities of the MLE approach are 3 students have too low ability and 6 students have the too high ability. Most students are at the level of ability $0 < \theta < 1$, and there are 52% of students above the average of mathematical literacy.

Item Information Function (IIF) represents information contributed by certain items that cross the range of abilities θ (Muraki, 1993). The item information curve shows how well and precisely each item can measure the latent trait across various levels of student ability (Figure 5). Item_1 and item_3 provide more information at a low level of mathematical literacy, item_2 provides more information at a moderate level of mathematical literacy, while item_4 can provide more information at a higher level of mathematical literacy.

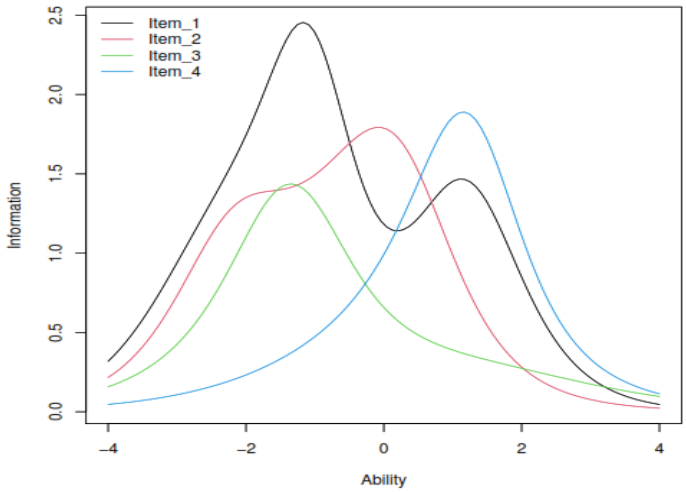

Figure 5. *Graph of item information function*

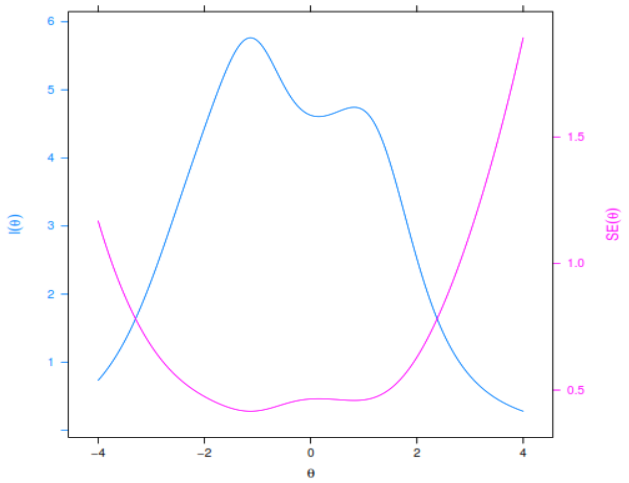

Figure 6. *Graph of the relationship between test ınformation function and standard error*

The sum of the information functions of the items making up the mathematical literacy test can be expressed as a test information function (Hambleton & Swaminathan, 1985). The test information function is correlated with the item information function. The test device information function will be high if the item has a high information function too. Item parameter index values and mathematical literacy capabilities of the analysis results are estimation results that cannot be separated from measurement errors. The standard error of measurement has a quadratic inverse relationship with the information function (Hambleton, Swaminathan, & Rogers, 1991). The relationship of the information function and standard error is presented in Figure 6. The mathematical literacy instrument is suitable for students with the ability $-2 < \theta < -0.5$ shown in the ability to have the maximum value for the information test or the minimum value for standard error.

CONCLUSIONS

Exploratory Factor Analysis results show that the mathematical literacy test measures 1 factor, so the analysis uses the unidimensional item response theory. The results of the model compatibility test show that GPCM is suitable for analyzing mathematical literacy test instruments with 4 levels of mathematical literacy. The estimation results of item parameters using GPCM indicate that level_1 item have a high item discrimination category, whereas level_2 item, level_3 item, and level_4 item have medium item discrimination categories. Based on the difficulty level of items, level_1 item, level_2 item, and level_3 item has easy categories, whereas level_4 item has difficult categories. Estimation of students' abilities uses 3 approaches, namely MLE, MAP, and EAP. Most students are at the level of ability $0 < \theta < 1$, and there are 52% of students above the average of mathematical literacy. Based on the test information function and the standard error shows that the mathematical literacy instrument is suitable for students with the ability $-2 < \theta < -0.5$ shown in the ability to have the maximum value for the information test or the minimum value for standard error.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, 13(2), 113-127. https://doi.org/10.1177/014662168901300201.

Afriyani, D., Sa'dijah, C., Subanji, S., & Muksar, M. (2019). Students' construction error in translation among mathematical representations. Journal of Physics: Conference Series 1157. https://doi.org/10.1088/1742-6596/1157/3/032098.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Brooks/Cole Publishing Company.

Baker, F. B. (1991). Comparison of minimum logit chi-square and Bayesian item parameter estimation. British Journal of Mathematical and Statistical Psychology. 44(2), 299–313. https://doi.org/10.1111/j.2044-8317.1991.tb00963.x.

Baker, F. B. (2001). Estimating an examinee's ability. In C. Boston, L. Rudner (Eds.), The basics of item response theory (pp. 85–102). ERIC Clearinghouse on Assessment and Evaluation.

Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. Australian Journal of Basic and Applied Sciences, 9(7), 549–556. https://doi.org/10.13140/RG.2.1.1561.5522.

Blum, W. & Ferri, R. B. (2009). Mathematical modelling: Can it be taught and learnt? Journal of Mathematical Modelling and Application, 1(1), 45–58. https://proxy.furb.br/ojs/index.php/modelling/article/view/1620/ 1087.

Blumberg, B., Cooper, D. R., & Schindler, P. S. (2005). Business research methods. McGrawHill Education.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459. https://doi.org/10.1007/BF02293801

Boughton, K.A., Klinger, D.A. & Gierl, M.J. (April 2001). Effect of random rater error on parameter recovery of the generalized partial credit model and graded response model. Paper presented at the annual meeting of the national council on measurement in education, Seattle, WA.

Brown, T.A. (2015). Confirmatory factor analysis for applied research (2nd ed.). New York: Guilford Press.

Camilli, G., Wang, M.-M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. Journal of Educational Measurement, 32(1), 79–96. https://doi.org/10.1111/j.1745-3984.1995.tb00457.x.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. IOSR Journal of Research & Method in Education, 3(1), 1–8. https://www.researchgate.net/publication/321268802.

Chang, L. (1994). A Psychometric evaluation of 4-poin and 6-point Likert-type scales in relation to reliability and validity. Apllied Psychological Measurement, 18(3), 205–215. https://doi.org/10.1177/014662169401800302.

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. Educational and Psychological Measurement, 70(5), 717–731. https://doi.org/10.1177/0013164410379322.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Harcourt Brace Jovanovich College Publishers.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. Measurement and Evaluation in Conseling and Development, 25, 172–189. https://psycnet.apa.org/record/1993-28125-001.

Doabler, C. T. & Fien, H. (2013). Explicit mathematics instruction: What teachers can do for teaching students with mathematics difficulties. Intervention in School and Clinic, 48(5), 276–285. https://doi.org/10.1177/1053451212473151.

Embretson, S. E., & Reise, S. (2000). Measuring persons: Scoring examinees with IRT models. In S. E. Embretson, & S. Reise (Eds.), Item response theory for psychologists (pp. 158–186). New Jersey: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58(3), 357–381. https://doi.org/10.1177/0013164498058003001.

Fiske, D. W. (2002). Validity for what? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 169–178). Mahwah, NJ: Lawrence Erlbaum.

Gick, M. L. (1986). Problem-solving strategies. Educational Psychologist, 21(1&2), 99–120. https://doi.org/10.1080/00461520.1986.9653026.

Gravemeijer, K., Stephan, M., Julie, C., Lin, F. L., & Ohtani, M. (2017). What mathematics education may prepare students for the society of the future? Int J of Sci and Math Educ, 15(Suppl 1), 105–123. https://doi.org/10.1007/s10763-017-9814-6.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. Psychometrika, 19, 149–161. https://doi.org/10.1007/BF02289162.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory. Boston, MA: Kluwer Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publication Inc.

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. Applied Psychological Measurement, 40(4), 289–301. https://doi.org/10.1177/0146621616631317.

Harrington, D. (2009). Confirmatory factor analysis. New York: Oxford University Press, Inc.

Jablonka, E. & Niss, M. (2014). Mathematical literacy. In S. Lerman, B. Sriraman, E. Jablonka, Y. Shimizu, M. Artigue, R. Even, R. Jorgensen, & M. Graven (Eds.), Encyclopedia of mathematics education (pp. 391–396). Dordrecht: Springer (Reference). Springer Science+Business Media.

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. Colombian Applied Linguistics Journal, 16(2), 167–184. http://dx.doi.org/10.14483/udistrital.jour.calj.2014.2.a03.
Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychology Measurement, 34, 111–117. https://doi.org/10.1177/001316446002000116.
Kang, T., & Chen, T. T. (2007). An investigation of the performance of the generalized S-X^2 item-fit index for polytomous IRT models. Research Report. ACT.
Kilpatrick, J., & Swafford, J. (2002). Helping children learn mathematics. Washington: National Academy Press.
Lawson, D. M. (2006). Applying the item response theory to classroom examinations. Journal of Manipulative and Physiological Therapeutics, 29(5), 393–397. https://doi.org/10.1016/j.jmpt.2006.04.006.
Linacre, J. M. (2006). Winstep: Rasch-model computer programs. Chicago: Winsteps.com.
Longford, N. (2008). Maximum likelihood estimation. In G. Casella, S. Fienberg, & I. Olkin (Eds.), Studying human populations: An advanced course in statistics (pp. 37–66). New York: Springer New York.
Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.
Maillard, O. A. (2019). Mathematics of statistiscal sequential decision making. Université de Lille Nord de France.
Mardapi, D. (2008). Teknik penyusunan instrument tes dan nontes [The technique of preparing test and nontest instruments]. Yogyakarta: Mitra Cendekia Press.
Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence centered design. Research Report. Princeton: Educational Testing Services. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159–176. https://doi.org/10.1177/014662169201600206.
Muraki, E. (1993). Information functions of the generalized partial credit model. Research Report. New Jersey: Educational Testing Service. https://doi.org/10.1177/014662169301700403.
Muraki, E. (1999). New appoaches to measurement. In G. N. Masters, & J. P. Keeves (Eds.), Advances in measurement in educational research and assessment. Amsterdam: Pergamon.
Myung, I. J. (2002). Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 47, 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7.
Naga, D. S. (1992). Pengantar teori skor [Introduction to score theory]. Jakarta: Gunadarma.
Nandakumar, R., Yu, F., & Zhang, Y. (2011). A comparison of bias correction adjustments for the DETECT procedure. Applied Psychological Measurement, 35(2), 127–144. https://doi.org/10.1177/0146621610376767.
Nickerson, R. (2011). Mathematical reasoning: Patterns, problems, conjectures, and proofs. New York: Psychology Press.
OECD (Organisation for Economic Co-operation and Development). (2004). A profile of student performance in mathematics: Learning for tomorrow's world – First results from PISA 2003.
OECD (Organisation for Economic Co-operation and Development). (2013). PISA 2012 assesment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. Paris: Author.
Österman, T. & Bråting, K. (2019). Dewey and mathematical practice: revisiting the distinction between procedural and conceptual knowledge. Journal of Curriculum Studies, 51(4), 457–470. https://doi.org/10.1080/00220272.2019.1594388.
Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer.
Resnik, M. D. (1997). Mathematics as a science of patterns. New York: Oxford University Press Inc.
Reys, R. E., Lindquist, M. M., Lambdin, D. V., & Smith, N. L. (2009). Helping children learn mathematics. USA: John Wiley & Sons, Inc.
Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. Applied Psychological Measurement, 32(5), 407–423. https://doi.org/10.1177/0146621607301278.
Schoenfeld, A. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), Handbook for research on mathematics teaching and learning (pp. 334–370). New York: MacMillan.

Stage, C. (2003). Classical test theory or item response theory: The swedish experience. Centro de Estudios Públicos, 42.
Stout, W. F. (2002). Psychometrics: From practice to theory and back (15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment). Psychometrika, 67(4), 485–518. https://doi.org/10.1007/BF02295128.
Tang, K.L. (1996). Polytomous item response theory (IRT) models and their aplications in large-scale testing program: Review of literature. Educational Testing Science. Princeton, NJ. RM-96-8 TOEFL Monograph Series.
Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), Large-scale assessment program for all students: validity, technical adequacy, and implementation (Edisi 1, pp. 181-211). Mahwah, NJ: Lawrence Erlbaum.
Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC: American Psychological Association.
Van de Walle, J., Karp, K., & Bay-Williams, J. (2010). Elementary and middle school mathematics methods: teaching developmentally (7th ed.). New York: Allyn and Bacon.
Van der Linden, W.J. & Hambleton, R.K. (1997). Item response theory: Brief history, common models and extentions. In W. J. Van der Linden, & R. K. Hambleton (Eds.), Handbook of modern item response theory. New York: Springer.
Ware Jr., J.E., Bjorner, J.B. & Kosinski, M. (2000). Practical implications of item response theory and compterized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. Medical Care, 38(9), II.73-II.82. https://www.researchgate.net/publication/12340368.