REID (Research and Evaluation in Education), 10(2), 2024, 214-226
Available online at: http://journal.uny.ac.id/index.php/reid

Comparison of item characteristic analysis models of reading literacy test with polytomous Item Response Theory

Firyal Nabila Harsana[1]*; Heri Retnawati[1]; Septinda Rima Dewanti[2]; Rogers Andrew Lumenyela[3]; Rimajon Sotlikova[4]; Maulana Fatahillah Adzima[5]; Atut Reni Septiana[6]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Islamic Queensland University of Technology, Australia
[3]Institute of Rural Development Planning, United Republic of Tanzania
[4]Webster University in Tashkent, Uzbekistan
[5]University of California Berkeley, United States
[6]University of Manchester, United Kingdom
*Corresponding Author. E-mail: firyalnabila.2020@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
19 August 2024
Revised:
27 September 2024
Accepted:
04 October 2024

Keywords
GPCM; Item Response Theory; reading literacy test; polytomous

ABSTRACT

This study aims to compare the analysis models of the characteristics of reading literacy items with the polytomous Item Response Theory, which uses the Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Nominal Reasons Model (NRM). This research is quantitative, whose secondary data were gathered from about 1,000 test takers' responses to reading literacy items in the 2018 reading literacy study analyzed with the R program. This model comparison was carried out so that the analysis results obtained were more accurate in representing the level of reading literacy in Indonesia. The results show that the GPCM model is the fit model with an AIC value of 23753.89 and a BIC value of 24042.45, and the number of suitable testlets is 7 out of a total of 7 testlets. Based on the relationship between information function scores and SEM, reading literacy items provide higher information when participants' abilities range between -2.3 and +2.

Scan Me:

To cite this article (in APA style):
Harsana, F., Retnawati, H., Dewanti, S., Lumenyela, R., Sotlikova, R., Adzima, M., & Septiana, A. (2024). Comparison of item characteristic analysis models of reading literacy test with polytomous Item Response Theory. *REID (Research and Evaluation in Education), 10*(2), 214-226.
doi:https://doi.org/10.21831/reid.v10i2.77852

INTRODUCTION

A test is one of the measuring instruments to collect information about a person's ability to respond to the questions tested. One example was given by the Language Development and Fostering Agency through the Center for Strategy Development and Language Diplomacy in the 2018 reading literacy study, which used a reading literacy test to reveal the level of reading literacy skills of students in Indonesia. The study was conducted as a technical policy preparation activity supporting national literacy activities and as a form of evaluation of the *Gerakan Literasi Sekolah* (GLS) program. This program was designed by the government because it refers to the PISA results in the field of reading ability from 2015 to 2022, which have decreased, with the scores sequentially being 397 in 2015, 371 in 2018, and 359 in 2022 (OECD, 2022).

ISSN: 2460-6995 (Online)

The reading literacy test consists of 40 items grouped into seven reading texts. The test was made into a reading test application to facilitate the process of collecting student reading literacy data. In the study, this test was used as an instrument because it is one of the simplest ways to measure the cognitive learning progress of test-takers.

To reveal the true ability of test takers, it is necessary to have a quality test instrument. Several methods can be used in analyzing test instruments, ranging from classical test theory and the Item Response Theory (Hambleton & Linden, 1982). Classical test theory is the basis of ability measurement by showing the relationship between observed test results and unobserved test results (Wang & Osterlind, 2013). Some aspects that need to be considered in the analysis include the item difficulty index, item discrimination index, the distribution of answer choices, and test score reliability (Safari, 2000; Mariati, 2009). Meanwhile, the Item Response Theory is a development of classical test theory and has three basic assumptions that must be met. The three basic assumptions include unidimensionality, local independence, and the accuracy of the item characteristic curve (Hambleton et al., 1991).

The Classical Test Theory has several limitations (Retnawati, 2016). In the Classical Test Theory, the individual's score obtained from a test is limited to the test being tested, so the test results are not possible to generalize beyond the test (Retnawati, 2016). Meanwhile, the Item Response Theory has several advantages over the Classical Test Theory. The advantages of the Item Response Theory, according to Bichi and Talib (2018), include (1) the ability of test takers does not depend on the level of item difficulty, (2) item statistics do not depend on the sample, (3) item analysis can use test items that are suitable for testing abilities at a certain level, (4) item analysis does not require strict parallel tests to estimate test reliability, and (5) item statistics and test-takers ability are both reported on the same scale.

In the Item Response Theory, in estimating students' ability to answer questions, two types of scoring can be used, namely dichotomous and polytomous. Dichotomous scoring is a scoring model for multiple-choice items, with the correct answer given a score of 1 and the wrong answer given a score of 0. The use of this scoring is often used because it is practical and easy, but a dichotomous scoring model cannot be used to distinguish errors made by test-takers because all wrong answers are scored 0.

Polytomous scoring is an assessment model with a wider scale of values; for example, the score on an item is between 0 and 2. According to Isgiyanto (2013), the polytomous scoring model can provide solutions to some of the limitations of the dichotomous scoring model related to measurement accuracy, the completeness of the attributes underlying the items, and the discovery of diagnostic information that has not been obtained from the dichotomous scoring model. Several models for polytomous scoring can be used, including the Rating Scale Model (RSM), Nominal Reasons Model (NRM), Graded Response Model (GRM), Partial Credit Model (PCM), and Generalized Partial Credit Model (GPCM) (Linden & Hambleton, 1997). To find out which model is most suitable for use in testing item analysis, it is necessary to test the suitability of the model. So that it can produce the actual ability of test-takers.

In the GRM, test takers' responses to items are categorized into ordered category scores, with the number of steps in correctly solving items (Retnawati, 2014). In this model, the relation-ship between item parameters and individual ability for the homogeneous case, in other words, when it is the same in each step, can be expressed as in Equation (1) and Equation (2) (Paek & Cole, 2020), where a_i is the item discrimination index of the item $i\,(i = 1,2, \dots, n)$ with a range of values $[0.1]$; θ is the level of an individual's ability with a range of values $[-\infty.\infty]$; b_{ik} is the difficulty index of the category k for the item i; $P_{ik}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of category k on the item i with a range of values $[-2.2]$; and $P^*_{ik}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of k or more on the item i.

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i(k+1)}(\theta) \quad \text{..... (1)}$$

$$P_{ik}(\theta) = \frac{\exp[a_i(\theta - b_{ik})]}{1+\exp[a_i(\theta - b_{ik})]} \quad \text{..... (2)}$$

PCM is a polytomous scoring model extending the Rasch model on dichotomous scoring (Retnawati, 2014). PCM assumes that each item's discrimination index is equal. In this model, the difficulty index in each step is not ordered, and one step can be more difficult than the next step. The PCM model by Paek and Cole (2020) is formulated as in Equation (3), where $P_{ik}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of category k on the item i; θ is the level of an individual's ability with a range of values $[-\infty. \infty]$; b_{ik} is the difficulty index of the category k for the item i with a range of values $[-2.2]$.

$$P_{ik}(\theta) = \frac{\exp \sum_{h=0}^{k}(\theta - b_{ih})}{\sum_{c=0}^{m} \exp \sum_{h=0}^{c}(\theta - b_{ih})}, k = 0,1,2, \dots, m \quad \text{..... (3)}$$

GPCM is an extended polytomous scoring model of PCM, in which the slope and discrimination parameters are estimated freely (Muraki, 1992). The general form of GPCM can be expressed in Equation (4) (Muraki, 1992), where $P_{ik}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of category k on the item i; a_i is item discrimination index of the item i $(i = 1,2, \dots, n)$ with a range of values $[0.1]$; θ is the level of an individual's ability with a range of values $[-\infty. \infty]$;b_{ik} is the difficulty index of the category k for the item i with a range of values $[-2.2]$; and D is the scale factor with a value of 1.7.

$$P_{ik}(\theta) = \frac{\exp \sum_{h=0}^{k}[Da_i(\theta - b_{ih})]}{\sum_{c=0}^{m} \exp \sum_{h=0}^{c}[Da_i(\theta - b_{ih})]} \quad \text{..... (4)}$$

NRM is a polytomous response model used when item responses are nominal (Bock, 1972). It can also be used for testlet modeling. The NRM is mathematically formulated as in Equation (5) (Paek & Cole, 2020), where $P_{ik}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of category k on the item i $(i = 1,2, \dots, n)$; a_{ik} is the item discrimination index of category k for the item i with a range of values $[0.1]$; θ is the level of an individual's ability with a range of values $[0.1]$; and b_{ik} is the difficulty index of the category k for item i with a range values $[-2.2]$.

$$P_{ik}(\theta) = \frac{\exp(a_i\theta - b_{ik})}{\sum_{h=1}^{m-1} \exp(a_i\theta - b_{ik})} \quad \text{..... (5)}$$

Many studies have used the polytomous scoring model, as has been done by Safitri (2020) regarding the comparison of estimating the mathematical literacy skills of grade VIII in Sragen City with the GRM, PCM, and GPCM models, which obtained the result that the GPCM model is the best model with nine items from 15 items matches and NFI of 14.936 and SEM of 0.259. In addition, the study also provided results that the overall mathematical literacy skills of grade VIII in Sragen City were in the moderate category. Bahar and Retnawati (2022), with their research on analyzing the characteristics of mathematical connection ability questions using GRM and GPCM model polytomous scoring, obtained the results that the GPCM model was the best model with three out of five items stated to fit the model. Santoso et al. (2022) examined the effect of scoring assessment and model fit on parameter ability estimation and participant fit on polytomous item response theory. The data used in the Santoso et al. (2022) research were the responses of 165 students in the Statistics course (SATS4410) to the test questions tested. The models used in item response theory analysis are PCM, GRM, and GPCM. The results obtained show that the GRM model is the best model fit based on the p-value and RMSEA. Their research further corroborates the advantages of employing sophisticated scoring models in educational assessments, demonstra-

ting that the choice of model can significantly influence the accuracy of ability estimations and the overall reliability of the assessment outcomes. Collectively, this research highlights the efficacy of polytomous scoring models, particularly the GPCM, in providing nuanced insights into students' mathematical literacy and capabilities, thereby underscoring the importance of selecting appropriate models for educational measurement.

In this study, we compared item characteristic analysis models of reading literacy tests with polytomous Item Response Theory to find out the characteristics of the items and the most suitable model for analyzing reading literacy tests and can represent the test takers' reading literacy skills most accurately and following reality. This study uses several models, including the Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Nominal Reasons Model (NRM). The rating scale model (RSM) model is not used because the data in this research are polytomous scoring with a different number of categories for each item. Unlike the studies described in the previous paragraph, this study compares model fit on four models and uses data from reading literacy tests.

METHOD

This research is quantitative. It describes Indonesian students' reading literacy skills and identifies the characteristics of reading literacy tests with the best polytomous model. The reading literacy test used in this study consists of 40 items in the form of multiple choice, essay, and description with seven reading texts (Language Development and Fostering Agency, 2018). The items were converted into a form of polytomous scoring with the testlet method, namely summing up the scores according to each reading text. Polytomous scoring is an assessment model with a wider scale of values; for example, the score on an item is between 0 and 2. Isgiyanto (2013) states that the polytomous scoring model can solve some of the problems of the dichotomous scoring model, including measurement accuracy, the completeness of the underlying characteristics of the items, and the discovery of diagnostic information that has not been found in the dichotomous scoring model. The reading literacy test was carried out by test-takers, namely grade X students of senior high schools throughout Indonesia, who were selected as research targets. The response results from the test individuals were then analyzed using the response theory of polytomous scoring items.

The data used in this study are secondary data, as many as 1,000 responses from the test set of reading literacy tests in the 2018 reading literacy study selected using the simple random sampling (SRS) technique. The data were obtained from research activities conducted by the Language Development and Fostering Agency in 2018. The research participants were anonymized to maintain confidentiality and guarantee that the obtained data were used only for research purposes.

The first step in data preparation was the selection process, with the result that two items were empty data. Thus, the data used in this study were 38 items, with details based on the reading text as follows: *Batik* (seven items), *Laskar Pelangi* (six items), *Penyakit Vektor* (eight items), *Perbandingan Musim* (six items), *Perpustakaan* (four items), *Drama* (four items), and International Program (three items). In the preparation, tests were mostly compiled based on the total score of a set of items (Thissen & Wainer, 2001). In line with the opinion of Lee et al. (2000) who state that testlets prioritize the unit of measurement rather than the grouping of items, the number of items in the testlet does not have to be the same. Then, the data were transformed using the testlet method based on seven reading texts with the help of R software (R Core Team, 2022). The results of the five data testlet methods are presented in Table 1. Once the data is shown in Table 1, it is ready for the polytomous item response theory analysis.

The Item Response Theory analysis in this study uses the Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Nominal Reasons Model (NRM). The four models are used in this study and can be used as a comparison because the data in this study are data with polytomous scoring which has a different number of categories on each item. The Rating Scale Model (RSM), which is also part of the polytomous scoring model,

is not used in this study because the model can only be used for models with polytomous scoring that have the same number of categories on each item. In testing the model's suitability, we paid attention to the number of items that matched each model with Yen's Q1 method, as well as the smallest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. Then, the ability parameters were estimated using the Bayes estimation method with the Expected A Posteriori (EAP) estimator.

Table 1. Results of Data Processing Ready to Use

Student	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6	Testlet 7
1	6	7	3	8	3	3	2
2	7	6	4	6	1	4	2
3	5	7	4	5	3	4	3
4	4	7	4	8	3	4	3
5	6	7	2	8	4	4	2

FINDINGS AND DISCUSSION

Findings

Comparison of the Fit of Polytomous IRT Models

The polytomous Item Response Theory models used in this study include GRM, PCM, GPCM, and NRM. Testing the suitability of this model was done statistically, which paid attention to the value of Yen's Q1 khi-squared generated in each model. The items could be suitable if the chi-squared values > chi-squared table or p-value > significance level, for which this study uses a significance level of 5% or 0.05. The results of the model fit test on the GRM, PCM, GPCM, and NRM models are presented in Table 2.

Table 2. Comparison of Model Fit Test Results

Testlet	GRM		PCM		GPCM		NRM	
	χ^2	p-value	χ^2	p-value	χ^2	p-value	χ^2	p-value
1	146.434	0.451	149.150	0.208	147.824	0.442	139.242	0.295
2	133.999	0.198	129.499	0.326	123.986	0.267	122.152	0.184
3	157.320	0.150	149.671	0.074	131.507	0.520	127.945	0.315
4	169.629	0.215	189.681	0.005*	164.104	0.237	149.388	0.298
5	78.214	0.440	83.337	0.348	63.050	0.738	76.883	0.355
6	68.228	0.571	88.861	0.283	58.289	0.860	57.088	0.846
7	85.988	0.094	85.566	0.190	82.930	0.091	72.029	0.204

Note. $*p < 0.05$, testlet does not fit the model

Table 2, shows that, of the four models used in this study, the number of testlets that fit the model the most is when using GRM, GPCM, and NRM modeling where all testlets fit. Meanwhile, for the PCM, one item that does not fit the model is the fourth testlet with a p-value <0.05. This, of course, cannot determine which model is the best because the number of items matched in the three models is the same. Thus, comparing by looking at other criteria, namely the AIC and BIC values in each model is necessary. The best model can be chosen based on the smallest AIC and BIC values. The results of the comparison of model fit based on the AIC and BIC values are presented in Table 3.

Table 3. Comparison of Model Fit Test Result Based on AIC and BIC

Model	Testlet Fit	AIC	BIC
GRM	7	23779.78	24069.33
PCM	6	23861.15	24121.26
GPCM	7	**23753.89**	**24043.45**
NRM	7	23770.40	24280.80

Based on the results in Table 3, the smallest AIC and BIC values are found in the GPCM model, where the GPCM model produces an AIC value of 23,753.89 and a BIC value of 24,043.45. Thus, the best model to use for analyzing data on reading literacy tests is the GPCM model.

Item Response Theory Assumption Test

Testing the assumptions of the Item Response Theory that need to be met includes the assumptions of unidimensionality, local independence, and parameter invariance. The unidimensional assumption means that each test item measures only one ability tested by factor analysis. The results of testing the unidimensional assumption are in the form of eigenvalues obtained for each test item and can be in the form of a scree plot presented in Figure 1.

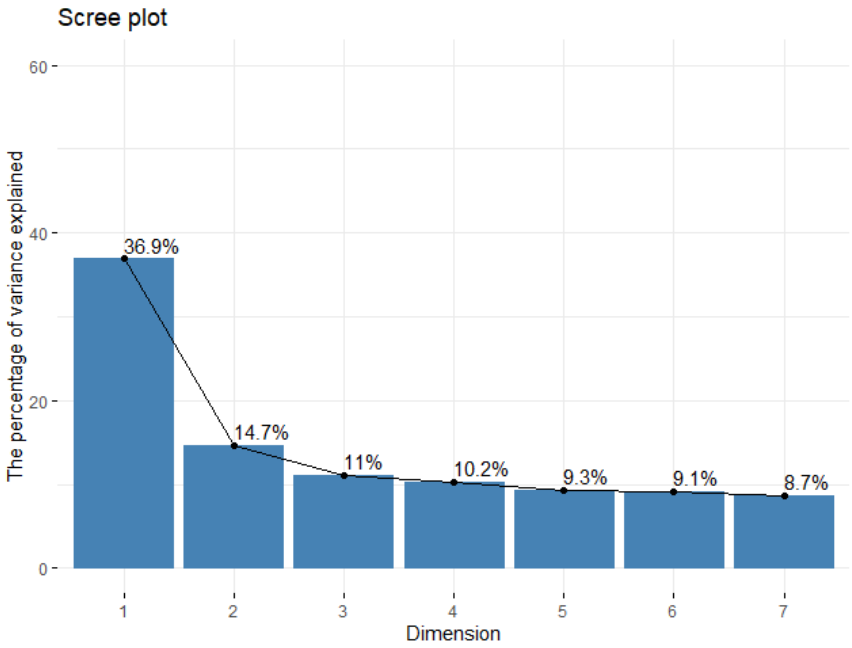


Figure 1. Scree Plot of Eigenvalues for Reading Literacy Test

Figure 1 shows one dominant component with a sharp steepness, and the others are sloping. In addition, the percentage of eigenvalue in dimension 1 is 36.9% or more than 20%. This supports the statement from Hambleton et al. (1991) that the first component with an eigenvalue of more than 20% is declared to fulfill the unidimensional assumption.

The local independence assumption test is a test to identify if the test taker's response to a test item does not affect the test taker's response to another item. The result of this test is in the form of inter-item correlation values on the reading literacy test presented in Table 4.

Table 4. Results of Correlation Values Between Items on the Reading Literacy Test

Testlet	1	2	3	4	5	6	7
1	1.00	0.29	0.35	0.19	0.25	0.30	0.16
2	0.29	1.00	0.24	0.30	0.32	0.38	0.32
3	0.35	0.24	1.00	0.17	0.21	0.22	0.11
4	0.19	0.30	0.17	1.00	0.24	0.25	0.30
5	0.25	0.32	0.21	0.24	1.00	0.30	0.28
6	0.30	0.38	0.22	0.25	0.30	1.00	0.31
7	0.16	0.32	0.11	0.30	0.28	0.31	1.00

Based on the correlation results in Table 4, it can be shown that the correlation value between items on the reading literacy test is less than 0.50, so it can be concluded that the assumption of local independence is met.

The parameter invariance assumption test is a test to prove whether the results of the item parameter estimation remain the same even though they are tested on groups of test takers with different ability levels. This test uses the best model, namely GPCM. The results of the invariance of item parameters in the form of discriminant index (a) and difficulty index (b) through the scatter diagram are shown in Figure 2(a) and Figure 2(b), respectively. Meanwhile, the ability parameter invariance results are shown with the scatter plot in Figure 2(c).

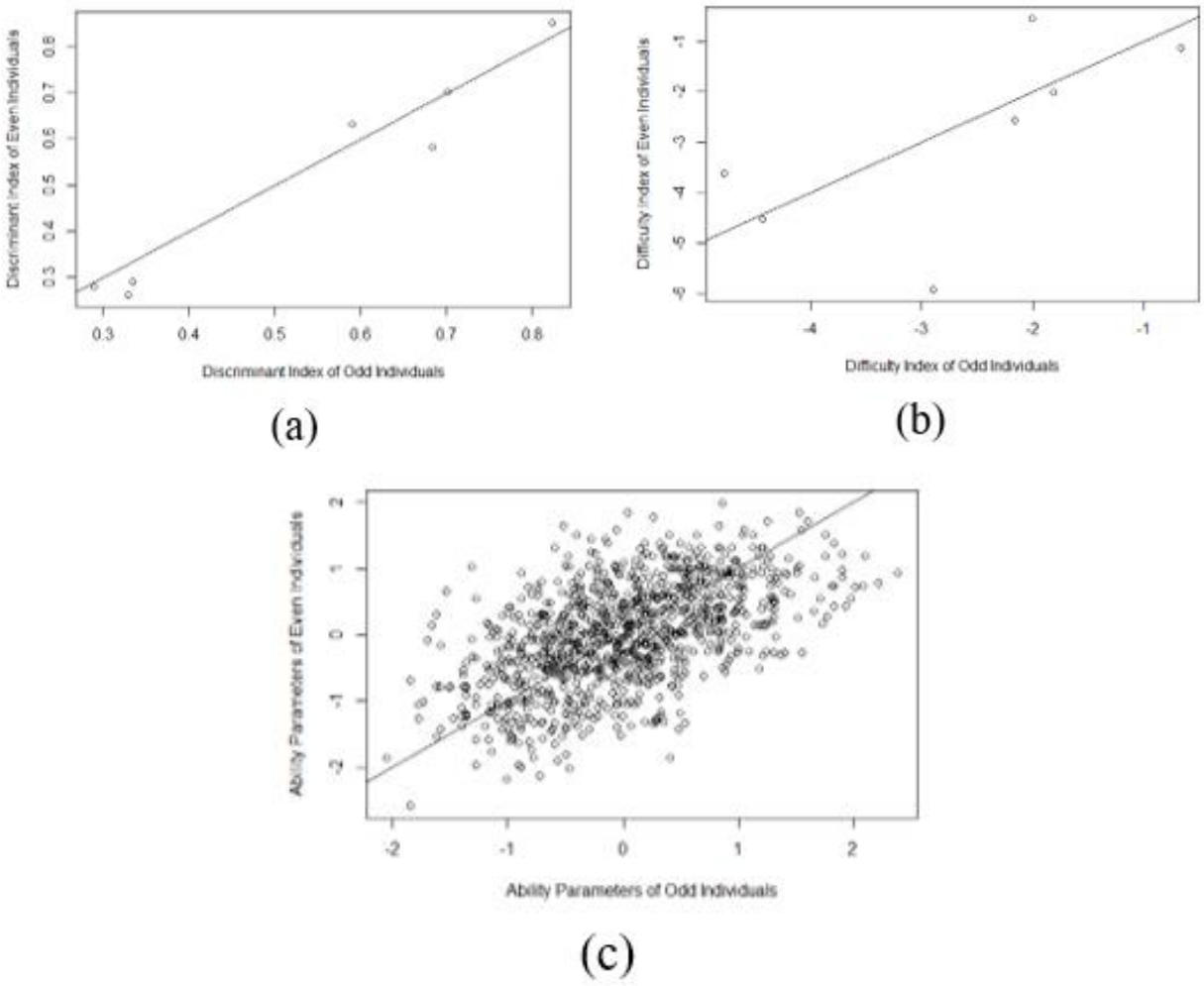


Figure 2. Scatter Plot of Parameter Invariance Assumption Test Results: (a) Discriminant Index; (b) Difficulty Index; (c) Ability Parameters

Based on these three figures, the points on the scatter plot are around the line and follow a straight line. Thus, it can be concluded that the assumption of invariance of item parameters and abilities in reading literacy test data is fulfilled..

Item Characteristics Analysis

This analysis was carried out by estimating the item parameters of a reading literacy test to obtain the discriminant index (a) and difficulty index (b) for the category parameters of each reading literacy test. The results of the item characteristics analysis on the reading literacy test are presented in Table 5.

Based on the results in Table 5, the grouping of criteria for the discriminant index, according to Sumarna (2006), shows that Testlet 1, Testlet 3, and Testlet 4 are considered sufficient because the value of a is between 0.21 and 0.40; Testlet 2 and Testlet 7 are considered good because the

value of a is between 0.41 and 0.70; Testlet 5 and Testlet 6 are considered very good because the value of a is between 0.71 and 1.00. Meanwhile, for the difficulty index, it can be seen that as the category of an item increases, the greater the difficulty level value becomes. Except for Testlet 1, Testlet 3 and Testlet 4, which have a discriminant index classified as sufficient, have a poor difficulty index because there is a decrease in value in the increasing category.

Table 5. Item Parameter Estimation Results with GPCM

Testlet	*a*	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11
1	0.31	-4.44	-4.05	-2.47	-0.52	-0.41	1.40	1.49	1.66	5.40	2.80	2.47
2	0.60	-4.13	-2.59	-2.20	-1.36	-0.28	0.19	1.49	2.94	4.76		
3	0.28	-4.46	-1.41	0.41	1.21	3.54	3.78	0.50	4.60	5.84		
4	0.29	-1.36	-3.50	-3.53	-2.60	-1.61	0.25	1.60	1.80	4.24	2.81	5.10
5	0.71	-1.19	-0.15	1.42	3.37							
6	0.84	-2.35	-1.45	-0.69	0.56							
7	0.63	-0.88	-0.05	1.84	4.52							

Information Function Values and SEM

The information function was used to indicate the amount of information that could be explained by each test item at various levels of latent properties. If the value of the information function was higher than the SEM, then the item on the test could be said to be good. The results of the test information function value and SEM of the reading literacy test data are shown through a curve in Figure 3.

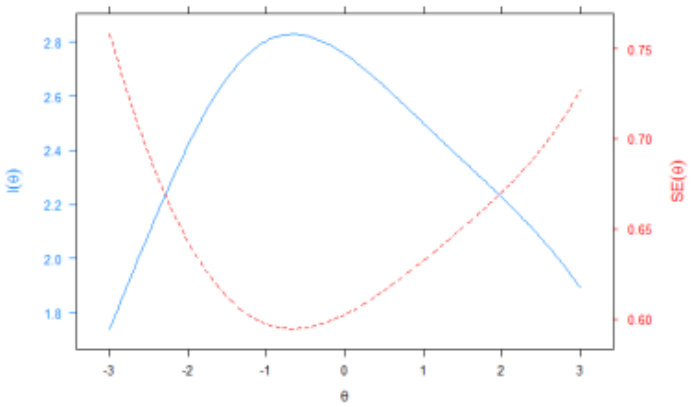


Figure 3. Information Function Values and SEM Curves

Figure 3 shows that there is a cut-off point resulting from the blue line in the form of the value of the information function and the red line in the form of standard measurement error (SEM). The cut-off point is on the ability scale (θ) between -2.3 and +2. At that interval, the reading literacy test instrument has a higher information function value than SEM. Meanwhile, if the reading literacy test is given to test takers with abilities outside the -2.3 to +2 range, they will provide a greater SEM value.

Indonesian Students' Reading Literacy Skills

The analysis of students' reading literacy ability was done by estimating the ability parameter (θ) with the best model obtained, namely GPCM. The estimation method used is the Bayes estimation method, where the estimator used is Expected A Posteriori (EAP). The descriptive statistics of students' reading literacy ability in the interval between -3 and +3 are presented in Table 6.

Table 6 shows that the average reading literacy ability of students in Indonesia is classified as moderate. This is shown in the value of mean ability with GPCM of 0.0009016, where the value is close to zero. This result is in line with research conducted by Hikamudin (2017) regarding the estimation of students' abilities in the national exam, which provides results that show that the

ability of SMA IPA students is mostly at a moderate level (average). Also, the overall reading literacy ability of students in Indonesia ranges from -2.5010513 to 2.1149332. In addition, the ability estimation results can also show the distribution of test takers' ability with GPCM, which can be seen in Figure 4.

Table 6. Descriptive Statistics of Students' Reading Literacy Skills

Statistics	Value
Mean	0.0009016
Maximum	2.1149332
Minimum	-2.5010513
Std. Dev.	0.8469477
Observation	1000

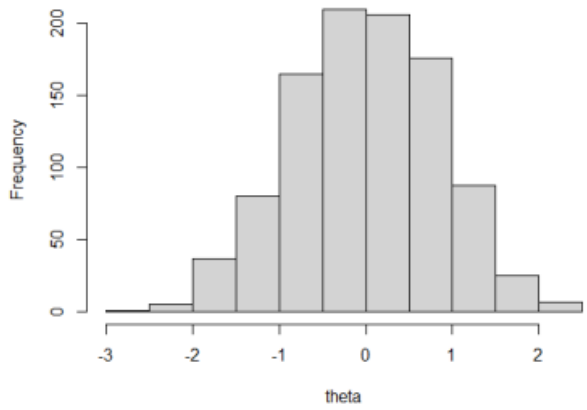


Figure 4. Histogram of Ability Estimation with GPCM

Figure 4 shows that the distribution of individuals' abilities in working on reading literacy tests with GPCM is mostly close to the mean value, around zero. To describe the distribution of test takers' reading literacy in more detail, it can be categorized based on the opinion of Manfaat and Anasa (2013), shown in Table 7. Table 7 shows that the distribution of students' reading literacy skills as a whole falls into the average category, with 755 out of 1000 individuals or 75.5%.

Table 7. Categories of Test Takers' Reading Literacy Skills

Interval Ability	Categories	Total	Percentage (%)
$-3.00 \leq \theta < -2.00$	Very low	6	0.6%
$-2.00 \leq \theta < -1.00$	Low	119	11.9%
$-1.00 \leq \theta < 1.01$	Average	755	75.5%
$1.01 \leq \theta < 2.01$	High	116	11.6%
$2.01 \leq \theta \leq 3.00$	Very high	4	0.4%
Total		1000	100%

Discussion

The reading literacy test, which consisted of 38 items and seven reading texts, was converted into testlet form so that the scoring on the test was polytomous (Susongko, 2010). One example of the reading text in the test set is *Penyakit Vektor*, and two examples of questions and their scoring guidelines are shown in Figure 5 and Figure 6.

Figure 5 shows an example of a text, Penyakit Vektor, a type of text used to send a letter of permission to be absent from school. In one text, there are several interrelated items. Two examples of items related to the *Penyakit Vektor* text are shown in Figure 6.

Surat izin tidak masuk sekolah salah satunya karena sakit dari orang tua atau wali murid saat ini masih diperlukan oleh pihak sekolah apalagi jika tidak terdapat surat keterangan sakit dari dokter. Surat ini dapat digunakan oleh sekolah sebagai rekaman atau arsip. Surat keterangan secara fisik dapat digunakan sebagai bukti autentik jika terjadi permasalahan yang memerlukan data dalam surat tersebut. Surat izin tidak masuk sekolah juga merupakan alat prosedural yang sesuai dengan aturan yang berlaku di sekolah sekolah. Surat izin tidak masuk sekolah dibuat berbasis kertas sehingga memerlukan sarana untuk mengirimkannya ke sekolah. Karena perlu pengantaran, surat izin tidak masuk sekolah juga memerlukan waktu untuk memprosesnya baik dalam penyusunan, pengiriman, maupun penyampaiannya. Surat juga bersifat fisik sehingga jika hilang tentu bukti autentik informasi tersebut juga dapat hilang.

Bandarlampung, 15 Agustus 2016

Yth. Bapak Drs. Suharjito
Wali kelas XII-MIPA
SMAN 2 Pringsewu-Lampung

Dengan hormat,
Saya beritahukan kepada Bapak bahwa anak saya,

Nama : Avelin
Siswa : Kelas XII-MIPA, SMA N 2 Pringsewu
Alamat : Jalan Mekarsari Nomor 404 Rejosari, Pringsewu

tidak dapat mengikuti pelajaran dan aktivitas di sekolah mulai hari ini, Senin, 15 Agustus 2016 sampai dengan beberapa hari ke depan karena sedang dirawat di Rumah Sakit Aminah. Anak saya terkena penyakit demam berdarah. Untuk itu, saya mohon Bapak memberikan izin kepada anak saya untuk beristirahat dalam pemulihan kesehatannya.

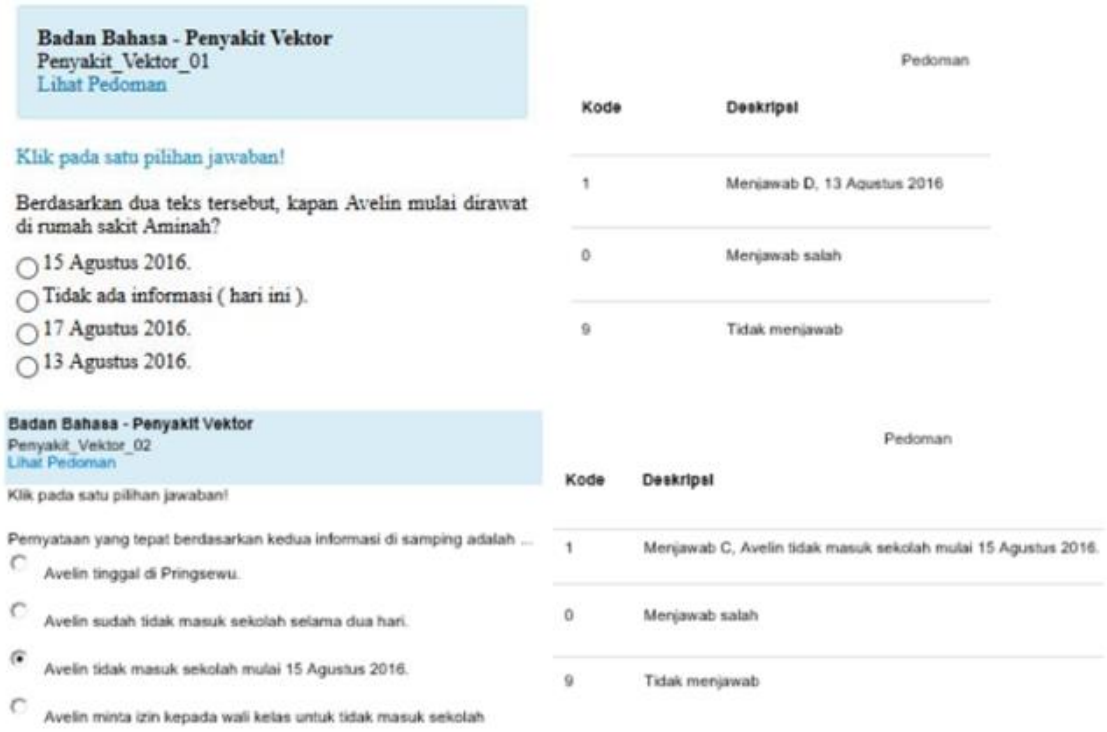
Atas perhatian Bapak, saya ucapkan terima kasih.

Hormat saya,
Orang tua

Oliver

Source: Language Development and Fostering Agency (2018)

Figure 5. Example of *Penyakit Vektor* Test



Source: Language Development and Fostering Agency (2018)

Figure 6. Test Questions and Scoring Guidelines

Figure 6 shows two related items since they use the same reading text, namely the reading text of *Penyakit Vektor*, as in Figure 5. In the scoring calculation, if the question is in the form of multiple choice, as in Figure 6, then each sub-item answered correctly is given a score of 1, and

incorrectly is given a score of 0. Meanwhile, for short-answer and description questions, score of 2 is given if the answer is correct and complete, a score of 1 is given if the answer is correct but incomplete, and a score of 0 is given if the answer is wrong. In addition, scoring on other reading texts can adjust to their respective guidelines.

The result of the model fit test using Yen's Q1 method, considering the AIC and BIC values in the GRM, PCM, GPCM, and NRM models, provides the result that the GPCM model is the best model to use in this study. This result is in line with the research conducted by Bahar and Retnawati (2022) regarding the analysis of the characteristics of the mathematical connection ability test of polytomous scoring between the GRM and GPCM models, with the result that the GPCM model is more suitable for use. However, the result of this study is not in line with research by Santoso et al. (2022) on the response data of test-takers in the Statistical Methods course, which gave the result that the PCM model was the best. This shows that the best model produced depends on the data used. Each datum has its characteristics, such as complex data structures, diverse response distributions, or different basic assumptions. In addition, different sample sizes and sample characteristics can also influence the selection of the best model.

Furthermore, based on the estimation of ability parameters carried out by the EAP method, the results show that the test individuals have reading literacy skills in the average category overall, where 75.5% of the total number of test takers had abilities in the range of -1.00 to 1.00. This result is in line with the research conducted by Ulpiyanti (2019), where the analysis of students' Mathematical critical thinking ability test with GPCM shows that most students have mathematical critical thinking ability with an average score of 61% of the total students. In addition, this result is also in line with research conducted by Safitri (2020), where the comparative analysis of the estimation of mathematical literacy skills of grade VIII in Sragen City shows that students' overall ability is in the medium or average category. It certainly shows that test takers' ability with the GPCM model's polytomous scoring is evenly distributed, with a medium average.

Average reading literacy results show that test-takers can complete reading tasks with a moderate level of complexity (Language Development and Fostering Agency, 2018). This is due to the lack of accessibility and quality of education in some areas, especially in rural areas. As stated by Tahmidaten and Krismanto (2020), the low reading literacy skills in Indonesia are partly due to the availability of reading materials, the implementation of learning activities, and the characteristics of practice or evaluation tests contained in teaching materials in schools that are still focused on low-order thinking skills. Therefore, improvement efforts are needed to improve the reading literacy of Indonesian students. For example, improving school library facilities and infrastructure, improving the quality of learning activities by teachers by applying methods such as SQR3 (Krismanto et al., 2015), Guided Practice (Boliti, 2017), Reciprocal Teaching Model (Noriasih, 2013), and others.

CONCLUSION

Based on the result of the study, it can be concluded that the GPCM model is the best of the four models under study, as indicated by the smallest AIC and BIC values, with an AIC value of 23753.89 and a BIC value of 24042.45. Also, the number of items that fit the model is seven out of a total of seven items. The result of the analysis of the characteristics of the reading literacy test with the GPCM model shows that the value of the discriminant and difficulty index is good, where the difficulty level of difficulty of an item will increase along with the increase in the category of each item. Based on the connection between the value of the information function and SEM, reading literacy tests provide higher information when individuals' abilities range between -2.3 and +2, with a maximum information function value of 2.83 at an ability of -0.7 and with a standard measurement error rate of 0.580. Based on the result of the ability parameter estimation, it can be concluded that the overall level of reading literacy ability of Indonesian students is in the average category, with 0.6% of test-takers in the very low category, 11.9% in the low category, 75.5% in the average category, 11.6% in the high category, and 0.4% in the very high category.

ACKNOWLEDGMENTS

The authors would like to express their gratitude and highest appreciation to Elisabeth Arti Wulandari from Clarkson University for her advice and valuable suggestions in improving this article.

DISCLOSURE STATEMENT

The authors declare that they have no conflicts of interest to disclose.

REFERENCES

Bahar, R., & Retnawati, H. (2022). Analisis karakteristik soal kemampuan koneksi matematika penskoran politomus. *Jurnal Tarbiyah,* 29(2), 195-211. http://dx.doi.org/10.30829/tar.v29i2.1650

Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education (IJERE), 7*(2), 142–151. http://doi.org/10.11591/ijere.v7i2.12900

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51. https://doi.org/10.1007/BF02291411

Boliti, S. (2017). Peningkatan kemampuan membaca pemahaman siswa kelas IV SDN 1 Lumbi-Lumbia melalui metode latihan terbimbing. *Jurnal Kreatif Tadulako, 2*(2), 12-23.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications Inc.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*(4), 373-378. https://doi.org/10.1177/014662168200600401

Hikamudin, E. (2017). Estimasi kemampuan siswa dalam ujian nasional menggunakan metode bayes. *Jurnal Penelitian Kebijakan Pendidikan, 10*(2), 1-14. https://doi.org/10.24832/jpkp.v10i2.171

Isgiyanto, A. (2013). Perbandingan penyekoran model rasch dan model partial credit pada matematika. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran, 43*(1), 9-18. https://journal.uny.ac.id/index.php/jk/article/view/1954

Krismanto, W., Halik, A., & Sayidiman, S. (2015). Meningkatkan kemampuan membaca pemahaman melalui metode survey, question, read, recite, review (SQ3R) pada siswa kelas IV SD Negeri 46 Parepare. *Publikasi Pendidikan: Jurnal Pemikiran, Penelitian dan Pengabdian Masyarakat Bidang Pendidikan, 5*(3). http://dx.doi.org/10.26858/publikan.v5i3.1616

Language Development and Fostering Agency. (2018). *Laporan kajian bahan kebijakan teknis literasi nasional tahun 2018*. Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15. http://dx.doi.org/10.1111/j.1745-3992.2000.tb00041.x

Manfaat, B., & Anasha, Z. Z. (2013). Analisis kemampuan berpikir kritis matematik siswa dengan menggunakan graded response models (grm). In *Seminar Nasional Matematika dan Pendidikan Matematika under the theme "Penguatan Peran Matematika dan Pendidikan Matematika untuk Indonesia yang Lebih Baik"* (pp. 119-124).

Mariati, I. (2009). Analisis butir soal dengan teori tes klasik (classical test theory) dan teori respons butir (item response theory) (Studi kasus: Soal ujian olimpiade sains provinsi bidang informatika 2009). *PYTHAGORAS Jurnal Pendidikan Matematika, 5*(2), 1-13. https://doi.org/10.21831/pg.v5i2.536

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychology Measurement, 16,* 159-176. https://doi.org/10.1177/014662169201600206

Noriasih, N. K. (2013). Pengaruh model pembelajaran reciprocal teaching terhadap pemahaman bacaan ditinjau dari konsep diri akademik siswa. *Jurnal IKA, 11*(2), 27-45. https://ejournal.undiksha.ac.id/index.php/IKA/article/view/1987

OECD. (2022). *PISA 2022 results: Factsheets.* OECD Publishing. https://www.oecd.org/en/publications/pisa-results-2022-volume-iii-factsheets_041a90f1-en/indonesia_a7090b49-en.html

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications.* Routledge.

R Core Team. (2022). *R: A language and environment for statistical computing.* https://www.R-project.org/

Retnawati, H. (2014). *Teori respons butir dan penerapannya (Untuk peneliti, praktisi, pengukuran, dan pengujian, mahasiswa pascasarjana).* Parama Publishing.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian.* Parama Publishing.

Safari. (2000). Panduan analisis butir soal. http://www.dikmenum.go.id/

Safitri, A. (2020). *Perbandingan estimasi kemampuan literasi matematika kelas VIII di Kota Sragen dengan berbagai model penskoran.* Master Thesis, Universitas Negeri Yogyakarta, Yogyakarta.

Santoso, A., Pardede, T., Djidu, H., Apino, E., Rafi, I., Rosyada, M. N., & Abd Hamid, H. S. (2022). The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory. *REID (Research and Evaluation in Education), 8*(2), 140-151. https://doi.org/10.21831/reid.v8i2.54429

Sumarna, S. (2006). *Analisis, validitas, reabilitas dan interpretasi hasil tes implementasi kurikulum 2004.* PT. Remaja Rosdakarya Offset.

Sumaryanta. (2021). *Teori tes klasik & teori respon butir: Konsep & contoh penerapannya.* Confident.

Susongko, P. (2010). Perbandingan keefektifan bentuk tes uraian dan teslet dengan penerapan graded response model (GRM). *Jurnal Penelitian dan Evaluasi Pendidikan, 14*(2), 269-288. https://doi.org/10.21831/pep.v14i2.1082

Tahmidaten, L., & Krismanto, W. (2020). Permasalahan budaya membaca di Indonesia (studi pustaka tentang problematika & solusinya). *Scholaria: Jurnal Pendidikan dan Kebudayaan, 10*(1), 22-33. https://doi.org/10.24246/j.js.2020.v10.i1.p22-33

Thissen, D., & Wainer, H. (2001). *Test scoring.* Routledge. https://doi.org/10.4324/9781410604729

Ulpiyanti, U. (2019). *Analisis tes kemampuan berpikir kritis matematis siswa dengan menggunakan generalized partial credeit model (GPCM): Penelitian deskriptif kuantitatif di SMP negeri 56 Bandung.* Doctoral Dissertation, UIN Sunan Gunung Djati Bandung, Bandung.

Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6

Wang, Z. & Osterlind S. J. (2013). Classical test theory. In T. Teo (Eds.), *Handbook of quantitative methods for educational research.* SensePublishers. https://doi.org/10.1007/978-94-6209-404-8_2