Multidisciplinary Output Research For Actual and International Issue (MORFAI Journal) ISSN (e): 2808-6635
Volumes 5 No. 4 (2025)

A MODERN APPROACH TO THE ACCURACY OF MATHEMATICAL PROBLEM-SOLVING ABILITY ASSESSMENT: GENERALIZED PARTIAL CREDIT MODEL

Aditya Prayogi[1]*, Mujahidawati[2], Ilham Falani[3]
[1,2,3]Megister Pendidikan Matematika Universitas Jambi, Indonesia
Corresponding E-mail: ilhamfalani@unja.ac.id*

Received	: 21 April 2025	Published	: 10 June 2025
Revised	: 30 April 2025	DOI	: https://doi.org/10.54443/morfai.v5i4.3190
Accepted	: 15 May 2025	Link Publish	: https://radjapublika.com/index.php/MORFAI/article/view/3190

Abstract

This study aims to measure students' mathematical problem-solving ability using the Item Response Theory (IRT) approach with *the Generalized Partial Credit Model* (GPCM). The research was carried out in North Bahar District, Muaro Jambi Regency, by involving all grade IX junior high school students in the even semester of the 2024/2025 school year as a sample through total sampling techniques. The test instruments were compiled and analyzed based on five main stages, including testing, scoring, and data processing and analysis using PARSCALE 4.1 software. The results of the analysis showed that the instruments used met the assumption of unidimensionality, which at the same time indicated the fulfillment of the assumption of local independence and parameter invariance. The model fit test produces a value indicating that the GPCM model matches the empirical data. Student ability estimates show a distribution that is close to normal, with most students being at moderate to slightly below average ability levels. The parameters of the question items showed high differentiating power and moderate difficulty, while the test information function showed the effectiveness of the instrument in measuring students' ability at average ability. In conclusion, the GPCM model is effectively used in measuring students' mathematical problem-solving abilities validly, accurately, and thoroughly.

Keywords: ***mathematical problem solving, Generalized Partial Credit Model, IRT***

INTRODUCTION

Mathematics is an important aspect of daily life, especially in improving human thinking. Therefore, mathematics is one of the subjects that is required at all school levels, from elementary school to the college level. In education itself, students' abilities are honed through problems so that students are able to increase their potential. Problem-solving is the process of overcoming the challenges that students face to achieve the expected goals(Kou et al., 2022). Problem-solving skills are a learning process that encourages students to actively participate in the learning process so that they can receive and respond well to questions and overcome problems and challenges that arise(Nisrina et al., 2021)Therefore, it is necessary to measure the student's ability to solve problems. This ensures that in practice students are able to maximize their problem-solving skills to face the challenges they will face in their daily lives.

Assessment tools should be prepared with careful attention to the material, construction, and language aspects, as this greatly affects student learning outcomes. Otherwise, the assessment tools or grades given by teachers will be inaccurate. Designing an assessment tool that functions as an instrument to test students' abilities requires the analysis of question items that have high validity and reliability, so that the distribution of easy, medium, and difficult questions is distributed proportionally according to the subject matter being tested(Zainal, 2020). In addition, good assessment skills assist teachers in identifying students' learning needs individually, allowing for timely and appropriate educational interventions. With effective assessment, teachers can ensure that each student gets the opportunity to develop optimally, as well as contribute to improving the overall quality of education(Scott, 2024).

Accuracy in assessing or measuring students' problem-solving abilities plays a very crucial role in the world of education. Proper assessments can provide an accurate picture of students' problem-solving abilities, which can ultimately assist educators in formulating more effective learning strategies(Zainal, 2020). The distribution and distribution of assessment results provides insight into the extent to which students understand the material being taught, so that it can be the basis for determining whether there is a need for improvement in the learning process.

The accuracy of the information obtained from the assessment is highly dependent on the instrument used. Therefore, the questions in the test must be calibrated to ensure their validity and reliability(Satria, 2024). Measurement is a process of determining numbers or quantifying individual characteristics or conditions based on certain rules. One of the commonly used methods in measurement is through the implementation of tests. There are two main forms of test instruments, namely description tests and objective tests(Prastiwi et al., 2023). The administration of this test is expected to produce accurate and precise measurement results. The accuracy in measuring students' creative thinking skills depends heavily on the quality of the test instruments used. This measurement is expected to be able to provide a clear picture of the extent of students' understanding of the subject matter, so that it can be used as a basis for evaluating and improving the learning process if necessary(Hayat, 2021).

In measurement, there are two main approaches that are often used to analyze question items, namely classical test theory (CTT) and item response theory (IRT)(Sarea & Ruslan, 2019). Classical test theory (CTT) is a basic theory in the measurement of mental ability that describes the relationship between the score observed on the test and the actual score that is not visible. CTT is group- and item-dependent, which means that the differentiating power index, difficulty level, and reliability coefficient of the test depend on who is taking the test and the question or item used(Retnawati et al., 2016). Research conducted by Hayat 2021; Fernanda and Hidayah 2020; Ciptari, Purwanti, and Erawati 2024 show that there are weaknesses that CTT has, namely that it is less effective in measuring the level of difficulty of question items and differentiating power compared to *modern test*. This weakness triggered a new and more adequate theory, namely *modern test* (modern test theory, also known as item/item response theory (TRA) or *item response theory* (IRT) and also known as *latent traits theory* (LTT). Although CTT has become a foundation, empirical interdependence and uniform reliability assumptions necessitated a shift towards more complex models such as IRT for better measurement accuracy (Frey, 2020).

Meanwhile, item response theory (IRT) is a general framework of mathematical functions that describes the interaction between individuals and question items(Sumintono & Widhiarso, 2014). IRT does not depend on a specific sample of questions or individuals taking the exam. One of the most popular IRT models, introduced by Georg Rasch in the 1960s was the Rash Model. This model is constantly evolving, not only for dichotomy analysis but also for polytomy data, one of which was developed by David Andrich from Australia(Kim & Wilson, 2020). One of the IRT Models for polytomy data is *Generalized Partial Credit Model* (GPCM). GPCM is also known as the Sorted Categorical Response Model because it deals with sorted polytomic categories, which can be related to constructed response items or selected responses, where test-takers are expected to get various score levels such as 0-4 points. In this case, the categories are 0, 1, 2, 3, and 4, which are sorted. 'Sorted' means that there is a certain order or rank of the response.

This study focuses on the accuracy of the assessment of mathematical problem-solving ability by using the IRT approach of the GPCM model for the analysis of polytomy data. The application of the GPCM model can increase the accuracy of the assessment of latent variables, which in this study is the mathematical problem-solving ability of students. This approach has several advantages that make it a good choice for test and questionnaire data analysis. One of its main advantages is its ability to predict lost data based on systematic response patterns(Kim & Wilson, 2020). This model is also suitable for analyzing data collected using scoring scales, Likert response scales, or other response data with sequential categories.

METHOD

This study uses an experimental quantitative approach and the data analyzed comes from students' responses to the mathematical problem-solving ability test, which was processed using PARSCALE 4.1 software and *Statistical Package for Social Science* (SPSS) software 26. The research was carried out in North Bahar District, Muaro Jambi Regency in January 2025 until it was completed. The population in this study is all grade IX junior high school students in the even semester of the 2024/2025 school year in the research area, consisting of 4 State Junior High Schools. The sampling technique uses total sampling, so that the entire population is used as a research sample.

The research procedure includes five main stages: (1) preparation of research instruments; (2) the implementation of test trials; (3) test result scoring; (4) data processing; and (5) data analysis, which includes testing model prerequisites, prerequisites for hypothesis testing, and research hypothesis testing. GPCM has a probability of scoring category k on the item. The GPCM model in calculating the estimated ability of participants takes into account the level of difficulty in each step. The GPCM model itself is similar to *the Partial Credit Model* (PCM). However, in the GPCM model, there are differentiating power parameters (a) and scale factor (D) which are the scale factor that has been set at 1.7. The mathematical model is as follows.

$$P_{jk}(\theta) = \frac{exp \sum_{v=0}^{k} Z_{jh}(\theta)}{\sum_{h}^{m} exp \sum_{v=0}^{k} Z_{jh}(\theta)}$$
$$k = 0,1,2,3, \dots, m$$

With the following equation:$Z_{jh}(\theta)$

$$Z_{jh}(\theta) = D_{aj}(\theta - b_{jh}) = D_{aj}(\theta - b_{jh} + d_h)$$
$$d_{j0} = 0$$

With

$P_{jk}(\theta)$ = probability of test takers with ability θ who obtain a category k score on point j.
aj = is the index of the difference in point j.
θ = Test taker ability
D = A scale factor of 1.7.
b_{jh} = Difficulty index in category k point to j.
b_j = Difficulty index at the location of point j.
d_h = Parameters in category k.

RESULTS AND DISCUSSION

RESULTS

Prerequisite Test

Before further data analysis was carried out using PARSCALE 4.1. There is a need to test the assumptions of IRT prerequisites. According to Embretson & Reise, (2000), unidimensional assumptions can be tested using factor analysis, with the help of (SPSS) 26. A one-dimensional test was carried out on the data before being used to estimate the parameters of the test participant's ability. The data tested are polytomy data for the GPCM model.

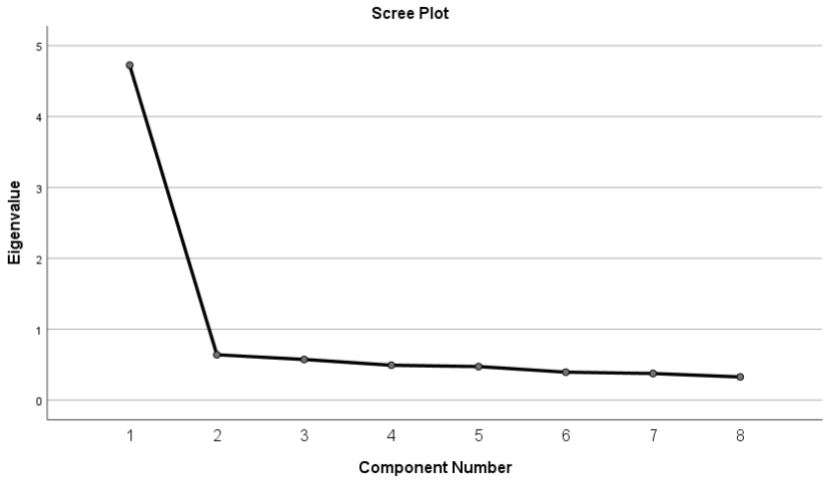
Table.1 Total Variance

Total Variance Explained

	Initial Eigenvalues			*Extraction Sums of Squared Loadings*		
Factor	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	4.724	59.055	59.055	4.724	59.055	59.055
2	.640	7.999	67.054			
3	.573	7.164	74.219			
4	.492	6.155	80.374			
5	.473	5.915	86.288			
6	.394	4.927	91.216			
7	.375	4.693	95.909			
8	.327	4.091	100.000			

Extraction Method: Principal Axis Factoring.

Based on the output of SPSS 26, the results of dimension reduction for the tested data showed that *principal axis factoring* extracted data into a number of factors with an eigenvalue of more than one. The data produced the main factor with a total *variance explained* of 59.05%, the second factor only contributed a total *variance explained* of 7.99%, while the rest had a contribution of total *variance explained* which ranged from less than 7.16%. The following Figure 1 is a graph of the scree plot data.



Gambar 1. *Scree Plot* Data GPCM

The results of the calculation of the factor analysis of the tested data and *the Scree Plot* in Figure 1. show that the main factors of each data can explain most of the total variance. So it can be concluded that the test items used are unidimensional. Most of the question items form a factor that can be called general math ability.

With the fulfillment of the one-dimensional test, the local independence test and the parameter invariance test are also considered to have been met, so it is enough to focus on the one-dimensional test only(Hambleton et al., 1991).

Model Fit Test (Person Fit)

Table 2. *Item Fit Statistics*

BLOCK	ITEM	CHI-SQUARE	D.F.	PROB.
BLOCK	0001	10.11094	10	0.431
BLOCK	0002	6.85348	10	0.740
BLOCK	0003	5.32487	10	0.869
BLOCK	0004	4.51313	11	0.952
BLOCK	0005	14.78008	10	0.140
BLOCK	0006	11.48368	10	0.321
BLOCK	0007	4.53074	10	0.920
BLOCK	0008	1.44694	10	0.999
TOTAL		59.04472	81	0.968

After the unidimensional prerequisite test is carried out, then a model fit test is carried out. Based on the results of the match test on the test item with the help of PARSCALE 4.1, *the item fit statistics* with $\mathcal{X}^2$ a GPCM value of 59.04 (p-value = 0.968) was obtained. Based on the *Item Fit Statistics* table above, all question items have a p-value (PROB.) greater than 0.05, with a total p-value of 0.968. This indicates that no item deviates significantly from the model, so it can be concluded that all items have a good match to the GPCM model. Thus, the question items in this instrument are valid to be used to accurately measure students' abilities.

Estimation Of Parameters Of Students'

Capability parameter estimation was carried out simultaneously and separately using the help of PARSCALE 4.1. To analyze student skills, it can be seen in Phase 3 on the PARSCALE output which is used to estimate students' skills. Table 3 presents a breakdown of each student's score, which illustrates the range of mathematical critical thinking skills between -2.41 and 2.38. This ability level was sorted from highest to lowest, based on the results of *a person measure* analysis of 108 students. In Table 3, *the value p* represents *the person*, while *m* represents *the measure*

Table 3. *person Measure*

p	*m*	*p*	*m*	*p*	*m*	*p*	*M*	*p*	*m*	*p*	*m*
105	2.31	74	1.16	48	0.43	108	-0.12	108	-0.43	27	-1.09
32	2.29	81	1.16	77	0.43	6	-0.14	6	-0.43	31	-1.09
73	2.15	90	1.16	86	0.38	11	-0.14	11	-0.44	36	-1.09
4	2.13	3	0.77	95	0.27	5	-0.15	5	-0.44	45	-1.09
64	1.9	10	0.77	23	0.24	52	-0.15	52	-0.61	62	-1.09
47	1.61	13	0.77	104	0.24	88	-0.16	88	-0.62	63	-1.09
55	1.58	85	0.77	39	0.23	57	-0.18	57	-0.62	9	-1.16
21	1.55	106	0.77	50	0.22	46	-0.19	46	-0.64	49	-1.47
1	1.29	28	0.67	92	0.21	103	-0.19	103	-0.64	99	-1.47
7	1.16	35	0.67	89	0.18	30	-0.35	30	-0.64	19	-1.5
8	1.16	53	0.66	97	0.17	35	-0.35	35	-0.64	94	-1.53
40	1.16	96	0.66	10	0.09	79	-0.37	79	-0.68	24	-1.54
54	1.16	58	0.47	33	0.07	72	-0.38	72	-0.71	44	-1.54
59	1.16	70	0.47	26	0.05	102	-0.38	102	-0.71	20	-1.79
65	1.16	75	0.47	51	0.02	16	-0.39	16	-0.71	78	-1.8
67	1.16	37	0.46	61	0.02	101	-0.39	101	-0.71	17	-1.82
68	1.16	80	0.45	41	0	82	-0.4	82	-1.09	15	-1.98
71	1.16	98	0.44	104	-0.02	100	-0.41	100	-1.09	14	-2.4

This assessment is carried out using *Person Measure* to estimate the level of ability of students based on their answers to each question item, which is expressed in logit units. The score shows the relative position of each student on an interval scale that is proportional to the difficulty level of the question, thus allowing for an objective comparison between students(Falani et al., 2017). Based on the analysis of 108 results of the students, it is known that the level of students' mathematical problem-solving ability is distributed sequentially with the details of each student's grades displayed in the following sequence distribution table.

Table 4. Distribution Frequency

No	Group	Median	Frequency	prob	Freq Komul	Prob Komul
1	-2,41 - -1,81	-2,11	3	3%	3	3%
2	-1,82 - -1,22	-1,52	8	7%	11	10%
3	-1,23 - -0,63	-0,93	18	17%	29	27%
4	-0,64 - -0,04	-0,34	25	23%	54	50%
5	-0,05 - 0,55	0,25	24	22%	78	72%
6	0,56 - 1,16	0,86	21	19%	99	92%
7	1,17 - 1,77	1,47	4	4%	103	95%
8	1,78 - 2,38	2,08	5	5%	108	100%

The table above presents the frequency distribution data of a variable divided into 8 classes of intervals, ranging from -2.41 to 2.38. Each class has a frequency, class median, probability, and cumulative frequency, as well as cumulative probability. Classes with a range of -0.64 to -0.04 had the highest frequency of 25, which is equivalent to 23% of the entire data. The next class that is also high is the interval of -0.05 to 0.55 with a frequency of 24 or about 22% of the total. These two classes (4th and 5th classes) account for about 45% of the total data, suggesting that most of the data is concentrated around near-zero values. The cumulative probability shows that up to the 5th class, it already covers 72% of the entire data, meaning that the majority of the values are below 0.55.The lowest class (interval -2.41 to -1.81) accounts for only 3%, and the highest class (1.78 to 2.38) accounts for 5% of the overall data. The following is a histogram of the estimated capability parameters.

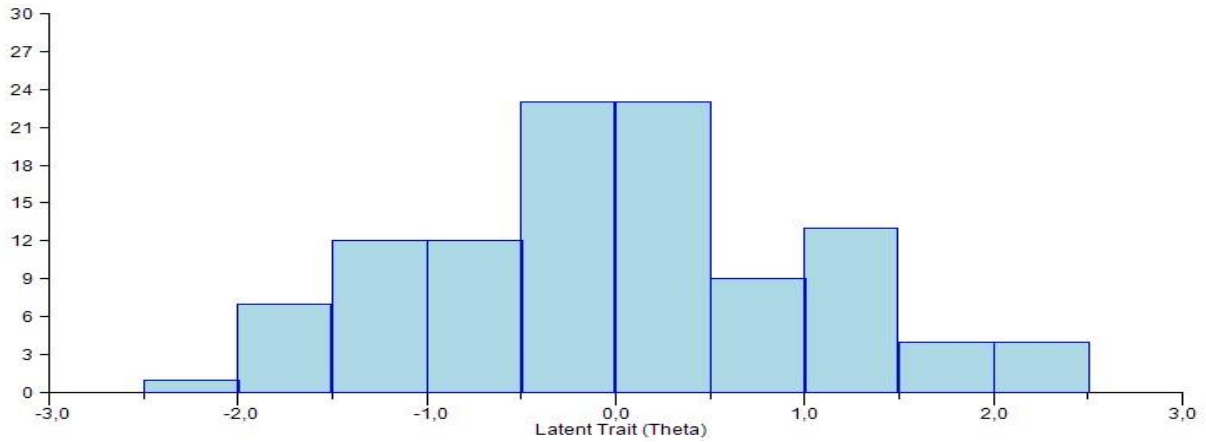


Figure 2. Histogram Estimation Of Parameter

The histogram above shows the distribution of latent trait values (θ) of participants based on the *Item Response Theory* model. The value θ represents the estimated level of ability of the participant, where a value of 0 indicates an average ability, a positive value indicates above-average ability, and a negative value indicates an ability below average. From the histogram, it can be seen that most of the participants had a θ value that was around 0 to -0.5, with the two highest bars indicating a frequency of about 24 participants in that ability range. This indicates that the majority of participants have abilities at a moderate or slightly below average level. The distribution appeared to be relatively symmetrical, although there was a slight left-skewed tendency, indicating that more participants were below average than those above average. Only a few participants had high (θ > 2) or very low (θ < -2) ability.

Results Of Estimating Item Parameters (*Item Measure*)

Table 5 *Item Measure*

ITEM	BLOCK	SLOPE	S.E.	LOCATION	S.E.	GUESSING	S.E.
0001	1	1.634	0.104	-0.001	0.036	0.000	0.000
0002	2	2.018	0.131	-0.045	0.032	0.000	0.000
0003	3	1.535	0.117	0.042	0.037	0.000	0.000
0002	4	1.560	0.114	0.071	0.039	0.000	0.000
0004	5	1.709	0.131	0.038	0.035	0.000	0.000
0005	6	1.896	0.128	0.133	0.034	0.000	0.000
0006	7	1.902	0.169	0.060	0.034	0.000	0.000
0008	8	2.106	0.152	-0.163	0.032	0.000	0.000

Based on the results of item parameter analysis using the *Item Response Theory model*, parameter estimates for 8 question items were obtained. The estimated parameters include the discrimination parameter (*slope*/a), the difficulty parameter (*location*/b), and the guessing parameter (*guessing*/c). In general, the discriminating value (a) for all items is in the range of 1.535 to 2.106, which indicates that all items have excellent discriminating power. A > 1.0 score indicates that an item is able to effectively distinguish between participants with low and high ability levels. The difficulty value (b) ranges from -0.163 to 0.113, which indicates that these items are on moderate difficulty. There are no very easy or very difficult items, so all items tend to be suitable for participants with average ability.Meanwhile, all items have a guess parameter value (c) of 0.000, which indicates that the model used is 2 Parameter Logistic Model (2PL), or guessing is not taken into account in the estimate. Thus, all question items can be said to be of good quality, with a high level of discrimination and a balanced level of difficulty. This supports the assumption that the instruments used can measure participants' abilities effectively and accurately.

Function Of Test Information

The test information function presents data related to the characteristics of a student's ability. There is an inverse relationship between the item information function and the standard measurement error (*Standard Error*

Measurement); This means that the smaller the measurement error rate, the higher the information that the test item can provide(Falani et al., 2020).

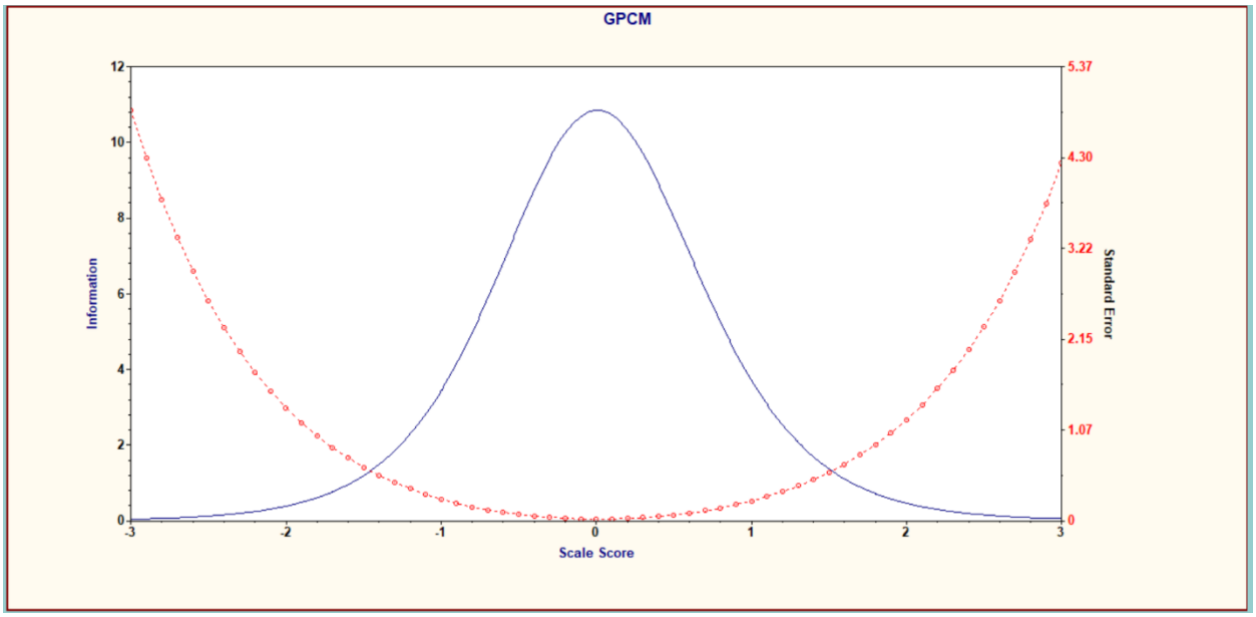


Figure 3. Test Information Function Graph

The figure above shows the standard information and error curves of a measurement instrument based on GPCM within the framework *IRT*. The blue curve describes the level of information provided by the test at different levels of ability (theta), while the red curve indicates the standard error value of the measurement. This test provides the highest information at theta values around -1, 0, and +1., which means this instrument is most accurate in measuring individuals with average ability. In contrast, at very low or very high theta values, the information provided decreases drastically, and standard errors increase, suggesting that measurements become less accurate in those areas.

DISCUSSION

IRT is a statistical approach used to describe the relationship between a person's latent ability and the response given to a test item. One of the models in IRT that is designed for data with tiered categories is GPCM. This model has several basic assumptions that must be met so that the results of the analysis do not contain bias. According to(Falani et al., 2017), the main purpose of the assumption testing in IRT is to ensure that the data has conformed to the basic principles underlying the model. In this study, three main assumptions namely unidimensionality, local independence, and parameter invariance have been met, so that the data are considered suitable for analysis using the GPCM model.

In addition to the assumption test, conducting a model fit test is an important stage in the implementation of IRT. This test aims to assess whether the applied model is in accordance with the empirical data obtained. When the model shows adequate match, then the estimation of parameters, such as the ability of the respondent and the characteristics of the item, can be performed accurately and reliably. The main purpose of the fit test in GPCM is to ensure the compatibility between the data collected and the designed model structure. The results of this study show that the instruments used are suitable to be analyzed using the GPCM model approach.

Based on the analysis of *the person measure score*, students' abilities are divided into three categories: high (*measure* > 1.5) as much as 7.92%, medium (0.5 ≤ *measure* ≤ 1.5) as much as 28.71%, and low (*measure* < 0.5) as much as 63.37%. These results show that students with low ability have a slightly larger proportion than the other two categories. In addition, all *LOCATION* values are around the number 0 with a very narrow range, which is between -0.163 to 0.134. This shows that the question items are best suited to measure students with average ability, because on the IRT scale, the theta value is close to 0.

CONCLUSION

This study shows that the measurement of students' mathematical problem-solving ability using the Item Response Theory (IRT) approach, especially the *Generalized Partial Credit Model* (GPCM) model, provides valid and accurate results. The instruments used have met the assumptions of unidimensionality, which indirectly also meet the assumptions of local independence and parameter invariance, making them suitable for further analysis. The GPCM model was proven to have a high compatibility with empirical data, which was shown by the results of the model fit test. The estimated results show that most students have abilities at a moderate to slightly below average level. In addition, the question items used in the instrument have high differentiating power and moderate difficulty, which makes them effective in distinguishing students' ability levels. Thus, the GPCM model is the right tool to measure students' abilities comprehensively and in-depth, and can be used as a reference in the development of assessment instruments in the future.

REFERENCES

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists multivariate*. London, UK: Erlbaum Publishers.

Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of Students' Ability Estimation on Combinations of Item Response Theory Models. *International Journal of Instruction*, *13*(4), 545–558.

Falani, I., Nisraeni, N., & Irdiyansyah, I. (2017). The ability of estimation stability and item parameter characteristics reviewed by Item Response Theory model. *International Conference on Education in Muslim Society (ICEMS 2017)*, 175–178.

Frey, F. (2020). Test Theory and Classical Test Theory. *The International Encyclopedia of Media Psychology*, 1–6. https://doi.org/10.1002/9781119011071.iemp0047

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hayat, B. (2021). Klasika: Program Analisis Item dan Tes dengan Pendekatan Klasik. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *10*(1), 1–11. https://doi.org/10.15408/jp3i.v10i1.20551

Kim, J., & Wilson, M. (2020). Polytomous item explanatory IRT models with random item effects: Concepts and an application. *Measurement*, *151*, 107062.

Kou, G., Yüksel, S., & Dinçer, H. (2022). Inventive problem-solving map of innovative carbon emission strategies for solar energy-based transportation investment projects. *Applied Energy*, *311*, 118680.

Nisrina, D., Simatupang, G. M., & Mujahidawati, D. (2021). *Pengaruh Model Problem Solving dalam Pembelajaran Jarak Jauh terhadap Kemampuan Pemecahan Masalah Matematis Siswa pada Materi Pola Bilangan di Kelas VIII MTs Negeri 5 Kota Jambi*. Universitas Jambi.

Prastiwi, Y. E. N., Al Barru, A. A., & Hidayatullah, A. S. (2023). Penilaian Dan Pengukuran Hasil Belajar Pada Peserta Didik Berbasis Analisis Psikologi. *Bersatu: Jurnal Pendidikan Bhinneka Tunggal Ika*, *1*(4), 218–231.

Retnawati, H., Hadi, S., & Nugraha, A. C. (2016). Vocational High School Teachers' Difficulties in Implementing the Assessment in Curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction*, *9*(1), 33–48.

Sarea, M. S., & Ruslan, R. (2019). KARAKTERISTIK BUTIR SOAL: CLASSICAL TEST THEORY VS ITEM RESPONSE THEORY? *Didaktika: Jurnal Kependidikan*, *13*(1), 1–16.

Satria, M. R. (2024). Transformasi Standar Penilaian Pendidikan Dan Revitalisasi Penilaian Pembelajaran Di Indonesia. *Jurnal Penelitian Kebijakan Pendidikan*, *17*(1), 57–66. https://doi.org/10.24832/jpkp.v17i1.930

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)*. Trim Komunikata Publishing House.

Zainal, N. F. (2020). Pengukuran, Assessment dan Evaluasi dalam Pembelajaran Matematika. *Laplace : Jurnal Pendidikan Matematika*, *3*(1), 8–26. https://doi.org/10.31537/laplace.v3i1.310