Advances in Social Science, Education and Humanities Research, volume 619

Proceedings of the 2nd International Conference on Innovation in Education and Pedagogy (ICIEP 2020)

Estimation of Test Item Parameters with Polytomous Item Response Using Partial Credit Model (PCM)

Suciati [1]*, Sudji Munadi[2], Sugiman[2]

[1]*Universitas Borneo Tarakan, Indonesia*
[2]*Yogyakarta State University, Indonesia*
**Corresponding author. Email: cauchy_my@yahoo.com*

ABSTRACT

Assessment activities would provide useful information if instruments developed and used have good quality. One of the standards for a quality test instrument is having acceptable item parameters. This research aims to analyze the quality of the test instrument in terms of the item parameters' quality with difficulty level on polytomous scoring. This research is explorative research by using a quantitative approach. The data used in this research is the test participant responses data to the items of numeracy. The participants have involved as many as 273 students of class VIII SMP in DI. Yogyakarta Province. The polytomous item response theory model used in this research is the Partial Credit Model (PCM). The estimation of item parameters is carried out using the R Program. The results showed that of the ten test items analyzed, only six items had the right criteria based on the item location parameters (level of difficulty of the stage) and Item Characteristic Curve (ICC). The information function value obtained is in the range -3.00 to 3.00, representing the instrument developed is suitable for test participant with abilities from -3.00 to 3.00 or from low abilities to high abilities.

Keywords: *Item Parameter, polytomous, PCM, R*

1. INTRODUCTION

An essential part of the learning process is assessment. If learning is likened to an adventure, then the assessment is a compass that determines the direction of adventure. Assessment is e process of collecting and unifying information in the form of attributes of an object based on measurement results, which are then interpreted based on a standardized standard [1],[2],[3],[4],[5],[6]. This information is gathered from tests and other measurements. The test results are then analyzed critically and integrated with additional information that will produce students' decisions.

Assessment activities will produce useful information if the instruments used are of good quality. The quality of the test items in an instrument needs to be analyzed to sort out the right and wrong items. Often, item selection uses only the Classical Theory Test (CTT), whose analysis is based on the index of difficulty and discrimination. Although the results were sometimes quite good, item analysis with CTT did not provide information on how test participants responded to different ability levels. One approach to test development that produces a complete picture is Item Response Theory (IRT).

CTT has several drawbacks which are enhanced by the IRT. The IRT tries to build a model of how latent psychological constructs can be expressed in ms of responses to observed items. In the classic test, the test participants' ability is estimated based on the correct answer. In item response theory, the ability is estimated by a non-linear function called a score. Such a model will provide beneficial information in developing, evaluating, and scoring tests.

The characteristics of IRT include: 1) Characteristics of the items do not depend on the test participants, 2) the score produced by the test participants does not rely solely on the test, 3) the model is expressed at the item level, not at the test level, 4) the model does not require rigorous parallel testing for the assessment of reliability, 5) the model provides a measure of precision for each ability [7],[8],[9]. The item response's mathematical model means that the subject's probability of answering the item correctly depends on the subject's ability and item characteristics. It means that test participants with high abilities will have a greater probability of answer correctly when compared to participants with low abilities. That is one of the advantages of IRT when compared to CTT.

Apart from the dichotomy model, another scoring model in the IRT, namely the polytomy item response model. In contrast to the dichotomy model, which is characterized by two categories of answer scores (0 and 1), the polytomy model has a response of more than two types where each step of the process is taken into account [10]. Polytomy item response models can be categorized into nominal and ordinal items, depending on data characteristics' assumptions. The nominal item response model can be applied to items that have alternative answers that are not ordered, and there are various levels of measured ability. Simultaneously, the ordinal response model occurs on items that can score into the number of specific categories arranged in the answer.

The polytomous models in the grain response theory include the N*ominal Response Model* (NRM), the Rating *Scale Model* (RSM), the P*artial Credit Model* (PCM), the *Graded Response Model* (GRM), and the G*eneralized Partial Credit Model* (GPCM) [11],[12]. Analysis of the items in this study will use PCM. PCM is a development of the Rasch Model 1-PL. The model is used to analyze test items that require several steps of completion. The item parameters analyzed only contained parameters for the item location/stage difficulty level. Operating characteristic functions/ OCF of PCM are defined by the following equation [13],[14]:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]}{\Sigma_{r=0}^{mi}\left[\exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]\right]} \qquad (1)$$

$P_{ix}(\theta)$ is The Probability of obtaining on item i, θ is the test participants' ability, and δ_{ij} is the Item location parameter j for item i.

The item location parameter on the PCM shows the point where two category probability lines meet one item. Location parameters of items at each stage do not have to be sequential [12]. It means that the difficulty level at one step can be more complicated than the difficulty level at a later step. Item location parameters will be easily understood by looking at the grain characteristic curve (ICC)/CRF graph. ICC describes the relationship between the chance to answer correctly($P_i\ (\theta)$)with the ability of the test participants (θ). ICC is a mathematical relationship related to the chance of success on the items measured by looking at the test's ability and the characteristics of the items [7]. The higher a person's ability, the more the chance to answer an item correctly will increase.

ICC's most famous mathematical form is the logistic form, whose graph is an S-shaped curve [15]. Absis at the ICC shows the test participants' ability, while the ordinate shows the chance to answer correctly. In general, above the intersection point items indicate relatively easy items because test participants with moderate ability can answer correctly with a chance of more than 0.5. On the other hand, the item whose curve is below the intersection point shows the chance to answer correctly is less than 0.5, which indicates the item is relatively difficult.

Item response theory has several assumptions that need to be verified before the modeling process is carried out. These assumptions are data unidimensionality, local independence, and invariance of measurement [16]. Unidimensionality indicates whether the model measures single or multiple attributes, local independence indicates whether responses to other items do not influence one item's response. In contrast, invariance shows that sample characteristics do not affect the model.

Unidimensional means that each test item measures only one ability [8]. In practice, the unidimensional assumption cannot be strictly fulfilled due to cognitive, personality, and test-taking factors, such as anxiety, motivation, and a tendency to guess. Therefore, the unidimensional assumption can be demonstrated if the test contains only one dominant component that measures subject achievement. Local independence is a condition in which the test participants' responses to any items will be statistically independent because other factors are constant. The concepts of unidimensional and local independence are interrelated. If unidimensional assumptions are met, then local independence will be fulfilled [17]. This assumption of local independence will be fulfilled if the participant's answer to one item does not affect the participant's response to another item.

The third assumption is parameter invariance. The parameter invariance concerning the sample's parameters' parameters to estimate model parameters, the parameter estimate will be linearly related to the parameter estimated with several other samples taken from the same population [18]. The test participant's abilities vary, from low to high. All test participant's low to high ability groups should invariably refer to the same item characteristic curve. The strength or contribution of test items in revealing the test's latent traits is expressed by the item information function [19]. Through the item information function, it can be seen which items are good and evil. The test information function is also the reliability of the IRT. The test information function is denoted by I (θ) and is estimated by the formula [14]:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad i = 1, 2, 3, \ldots n \qquad (2)$$

I (θ) is the Information function test, and I_i (θ) is information function item i.

Based on the mathematical equation of the information function items and the test information function, the test information function is a linear combination of the item information function. Therefore the test information function will have a high value if the item information function value is also high. Lord [20] expresses the function item information with the following equation:

$$I_i\,(\theta) = \frac{[P_i'\,(\theta)]^2}{P_i\,(\theta)Q_i\,(\theta)} \qquad i = 1, 2, \ldots n \qquad (3)$$

Where I_i (θ) is the information function item i, $P_i\,(\theta)$ is the probability of the participant's ability to answer correctly θ point i. $P_i\,'(\theta)$ is derivative function $P_i\,(\theta)$ against θ, and $Q_i\,(\theta)$ is odds participants with the ability to θ answer correctly point i. Test information function relating to the measurement error. The amount of information generated from a test on a specific ability is inversely proportional to the estimated ability precision and is expressed by the mathematical equation [21]:

$$\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{I\,(\theta)}} \qquad (4)$$

S.E. $(\hat{\theta})$ is the estimation error, and I (θ) is Test Information Function. The test information function is inversely proportional to the measurement error. It means that the higher the value of the information function, the smaller the error in estimating the test participants' ability. Therefore, a good test kit has a high information function because it shows accuracy and precision in estimating the ability of test participants. This study aims to reveal the characteristics of test items that have been tested and have evidence of validity.

2. METHODS

The type of research used to reveal the test items' characteristics is exploratory research with a quantitative descriptive approach. The test items to be analyzed are Mathematical Literacy test items developed by the researcher and have proof of the content and construct validity. The empirical data used comes from the mathematical literacy instrument's test results with a total sample of 273 students of class VIII SMP in DI. Yogyakarta Province. There were ten items analyzed using polytomous scoring. The questions given are developed systematically in the form of story questions that use narration and pictures to describe the situation of the problems raised.

After working on the question items, the test participant's response data were analyzed using the IRT modeling procedure for polytomous scores. The technique used is testing the IRT assumptions, which are then continued to estimate the item parameters with the *Partial Credit Model* (PCM) using the R application.

Confirmation of the unidimensional assumptions is carried out with SPSS assisted factor analysis. The unidimensional assumption test is carried out by looking at the eigenvalues and the percentage of variance in factor analysis. Based on the eigenvalues' magnitude, two criteria become the conditions for fulfilling the unidimensional assumptions, namely 1) the first eigenvalues are much greater than the second eigenvalues, 2) the second eigenvalues are not too big compared to the other eigenvalues [24]. Based on the output of factor analysis and *scree-plotEigenvalues and the percentage of variance.* The fulfillment of the unidimensional assumption also shows the fulfillment of the local independence assumption.

After testing the IRT assumptions, the next step is to estimate the item parameters. Parameter estimation was done using PCM. The parameter estimated on the PCM is the level of grain difficulty. The item difficulty level is in the interval -∞ to +∞. Good grains have a difficulty level between -2 to +2 [7]. The final step taken is to estimate the information function of the test.

3. RESULT

After working on ten items, responses to the test participants scored Mathematical Literacy using the developed scoring guidelines. Scoring is done by looking at the stages of students solving questions based on a combination of criteria. Score 2 (*full credit*) if all phases are correct, score 1 (*partial credit*) if only part of the steps is correct, and score 0 (*no credit*) if there are no correct steps. Item characteristics are estimated using the PCM polytomous model approach with *software* R. Before estimating item parameters, it is necessary to confirm the item response theory's assumptions. The assumptions are data unidimensionality, local independence, and parameter invariance.

Unidimensionality indicates whether the model measures single or multiple attributes. Confirmation of assumptions is carried out using factor analysis with the help *of SPSS software.* The analysis begins by checking the adequacy of the sample. Analysis of the adequacy of the sample using the KMO and Bartlett test results. The results of the analysis are presented in Table 1.

Table 1. KMO and Bartlett Test Results

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0,781
Bartlett's test of sphericity	*Approx Chi-Square*	608,608
	df	45
	significance	0,000

Based on Table 1, the *Chi-Square value* sample in the Bartlet test is 608.608 with 45 degrees of freedom, value *p* <0.05, and a KMO value of 0.781. A group of data is said to meet the sample assumption's adequacy if the KMO value is> 0.5 [22]. It means that the sample size of 273 used in the instrument trial was sufficient; therefore, what can continue the analysis. The analysis was carried out by looking at the number of eigenvalues and variance in the total output *variance* from the factor analysis results. The recapitulation of the analysis results is presented in Table 2.

The analysis shows that the ten items analyzed are grouped into three main factors: the number of components that have eigenvalues ≥ 1. Percentage of variance explained from each of these factors. The first factor has a variant of 33,822%, the second factor is 12,423%, and the third factor is 10,864ined %, so that these three factors can explain about 57.109% of the total variant.

Table 2. Eigenvalues and Proportions of Variants

Component	Eigen Value	Percentage of Variance	Percentage Cumulative of Variance
1	3,382	33,822	33,822
2	1,242	12,423	46,245
3	1,086	10,864	57,109
4	0,911		
5	0,854		
6	0,667		
7	0,563		
8	0,510		
9	0,438		
10	0,346		

Variance percentage shows that the first component is more dominant because it has a much larger percentage of the variance than the other two components. Eigenvalues and the percentage of variance can also determine the number of factors formed by observing the *Scree plot* of the distribution of eigenvalues. *Figure 1 shows the Scree plot of* eigenvalues.

Based on the *scree plot* in Figure*1,* there is one slope resulting from changes in the decline of eigenvalues in the first and second components. The steepness occurs because the first and second components' eigenvalues are very different as for eigenvalues of the second component to tenth component *slope* generated inclined ramps. It is because the eigenvalues of the second component and the other components are not much different. It means that the first component is very dominant compared to other components. In conclusion, the model measures a single attribute because there is only one measured dimension in the instrument. Unidimensional assumptions are met in the Mathematical Literacy test instrument developed.

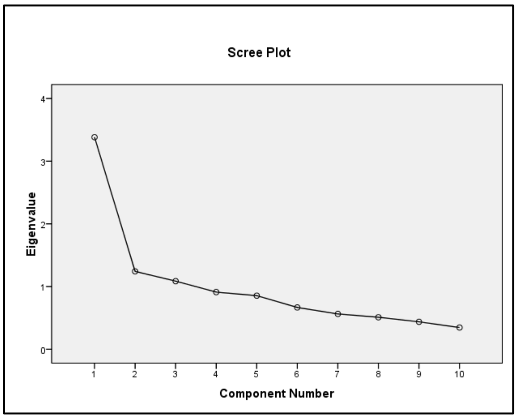

Figure 1. ***Scree Plot*** **Factor Analysis**

The fulfillment of the unidimensional assumption also proves the local independence assumption's completion. The concepts of unidimensional and local independence are interrelated. The unidimensional concept shows that the instrument developed measures only one dimension, that is, Mathematical Literacy. Local independence is a condition in which the subject's response to any pair of items will be statistically independent because factors that influence performance are constant.

After testing the IRT assumptions are fulfilled, the next step is to estimate the parameters of the Mathematical Literacy test items being developed. Item parameter estimation was carried out to determine the quality of the developed questions. The item parameter estimated was the grain difficulty level. Results of analysis for polytomous scoring with PCM obtained an estimate of the difficulty level of the developed ten mathematics literacy test items. The level of difficulty on the PCM consists of the stage difficulty level and the item difficulty level. The analysis of the test item parameter estimates in the field trial is presented in Table 3.

The level of difficulty of the step is a transition from one category to another. The difficulty level is the intersection of the item characteristic curve between two consecutive types having the same chance of being owned by test participants with specific abilities. Mathematical Literacy items have a polytomous scoring with three categories. Therefore there are two steps of difficulty. The level of difficulty of step 1 is the intersection of the grain characteristic curve between categories 1 and 2. Step 2 is the intersection of the item characteristic curve between categories 2 and 3. The ten items' estimation results' stage difficulty levels are in the range of –1,534 to 1,933.

Table 3. Estimation of Parameters for Mathematical Literacy Test Items

Item number	Difficulty	Level of Difficulty		Decision
		Step 1	Step 2	
1	-1,0925	-1,385	-0,8	Good
2	0,8375	-0,258	1,933	Good
3	-0,3925	-1,534	0,749	Good
4	1,18	1,936	0,424	Good
5	0,161	0,192	0,13	Good
6	0,815	0,444	1,186	Good
7	-0,4085	-0,478	-0,339	Good
8	-0,293	-0,204	-0,382	Good
9	-0,1415	-0,451	0,168	Good
10	0,6765	0,446	0,907	Good

The item difficulty level is the average of the steep difficulty level. The result of the item difficulty level calculation shows that the item difficulty level is in the range of -1.0925 to 1.18. It means that the difficulty level of the ten items developed is in the excellent category. It refers to the opinion of Hambleton et al. (1991: 13), which states that useful items have difficulty levels between -2 to +2.

The item difficulty level is proportional to the test participants' ability. The test participants' ability is relative to the items. According to Keeves and Alagumaia, if the test participant’s ability exceeds the item difficulty level, then the test response is expected to be correct. If the test participant's ability is less than the item difficulty level, then the test participants' response is expected to be wrong. Item location parameters will be easy to understand by looking at the ICC. ICC results of the analysis with the help of the R program are presented in Figure 2.

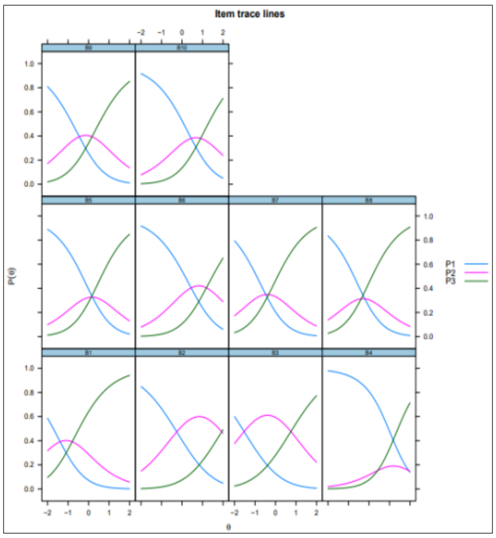

Figure 2. Item Characteristic Curve

Figure 2 presents the relationship between the probability to answer correctly(P_i (θ))with the ability of the test participant's (θ) of each item. Items 1, 2, 3, 6, 9, and 10 have a relatively good ICC because each category can distinguish the test participants' abilities. The ICC was not well demonstrated by ICC item number 4 because it had an irregular pattern. The categories were unable to distinguish the abilities of the test participants.

The relationship between the test and the individual being measured will produce accurate measurement information. The accuracy of the measurement results on various parameters of ability (θ) is indicated by the item information function's estimation results or the test. Figure 3 presents a graph of the information function and *standard error* of the Mathematical Literacy test produced by the R program.

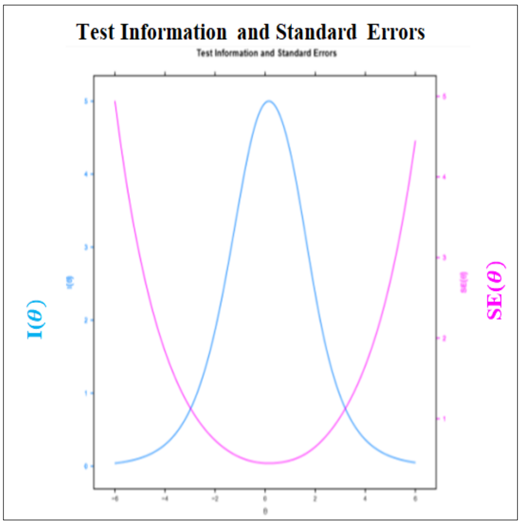

Figure 3. Test Information and Standard Errors

The pink curve gives the *standard error of* the measurement results, and the blue curve shows the value of the mathematics literacy test's information function. Based on its constituent items' information function, Mathematical Literacy instruments have a higher information function value than their measurement error. The value of the information function obtained is in the range of -3.00 to 3.00. The value of the maximum information function received based on the graph is 5. The instrument developed is suitable for test participants with abilities from -3.00 to 3.00 or test participants with low to high abilities.

4. CONCLUSION

The results showed that of the ten test items analyzed, all had an adequate level of difficulty. The value of the information function obtained is in the range of -3.00 to 3.00, and an instrument developed suitable for test participants with abilities from -3.00 to 3.00 or from low to high abilities. Even though what fulfilled the test of the IRT assumptions and all items were of good quality, PCM's use did not seem appropriate. The ICC curve of the four items shows an irregular pattern. Therefore it is necessary to test the fit of the model. The item analysis approach with PCM may not be suitable for the instrument being developed. Analysis using other models can be carried out for improvement.

REFERENCES

[1] Anderson L. W, *Classroom assessment: Enhancing the quality of teacher decision making*, New Jersey: Taylor & Francis e-Library, 2008

[2] Mardhapi D, *Pengukuran Penilaian dan Evaluasi Pendidikan* (2nd ed.), Yogyakarta: Parama Publishing, 2017.

[3] Nitko A. J & Brookhart S. M, *Educational assessment of Students* (5th ed.), New Jersey: Pearson Education, Inc, 2007.

[4] AERA, APA, & NCME, *Standards For Educational And Psychological Testing*. Washington, DC: American Educational Research Association, 2014

[5] Miller P., *Measurement and teaching*. USA: Patrick W. Miller & Associates. 2008.

[6] Burton M., Silver E A., Mills V L., Audrict W., Strutchens M. E., & Petit, M, “Formative assessment and mathematics teaching: Leveraging powerful linkages in the U.S. context”, *Classroom Assessment in Mathematics: Perspectives from Around the Globe*, Cham: Springer International Publishing, 2018, pp. 193-205,

[7] Hambleton R. K., Swaminathan H., & Rogers H. J, *Fundamentals of item response theory* (Vol. 2), Newbury Park, CA: Sage Publication Inc, 1991.

[8] Azwar S., *Dasar-Dasar Psikometri (Edisi II)*, Yogyakarta: Pustaka Pelajar, 2016.

[9] Keeves J.P., & Alagumalai S., New approaches to measurement, *Advances In Measurement In Educational Research And Assessment*, Amsterdam: Pergamon. 1999.

[10] Lee H.Y., & Dodd B.G, Comparison of Exposure Controls, Item Pool Characteristics, and Population Distributions for CAT Using the Partial Credit Model, *Educational and Psychological Measurement*, Vol.72, no.1, pp. 159-175, 2012.

[11] DeMars C., *Item Response Theory*, New York: Oxford University Press, Inc, 2010.

[12] Retnawati H., *Teori Respon Butir dan Penerapannya*, Yogyakarta: Parama Publishing, 2014.

[13] Engelhard G., Item Response Theory (IRT) Models for Rating Scale Data, *Encyclopedia on Statistics in Behavioral Science*, https://doi.org/10.1002/9781118445112.stat06398, 2014, pp. 995–1003.

[14] Desjardins C.D., & Bulut, O, *Handbook Of Educational Measurement And Psychometrics Using R*, New York: CRC Press, 2018.

[15] Primi C., Morsyanyi K., Chiesi F., Donati M. A., & Hamilton, J., The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT), *Journal of Behavioral Decision Making*, Published online in Wiley Online Library (wileyonlinelibrary.com) DOI:10.1002/ bdm.1883, 2015.

[16] Crocker L., & Algina J., *Introduction to classical and modern test theory*, New York: Holt Rinehard And Winston Inc, 1986.

[17] Embretson, Susan E., Reise, & Steven Paul, *Item Response Theory for Psychologists Multivariate Applications Book Series*, New Jersey:Lawrence Erlbaum Associates, 2000.

[18] Bejar Isaac I., Introduction to Item Response Models and Their Assumptions, In Hambleton R. K (*Eds*), *Applications of Item Response Theory*, Canada: Educational Research Institute of British Columbia,1983.

[19] Istiyono E., *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika*, Yogyakarta: UNY Press, 2018.

[20] Lord F. M., *Applications of item response theory to practical testing problems*, New Jersey: Lawrence Erlbaum Associates Publisher, 1980

[21] Widhiarso Wahyu, *Model Politomi Dalam Teori Respons Butir*, Available at SSRN: https://ssrn.com/abstract=2593459, 2010.

[22] Hair J. F., Black W. C., Babin B. J., & Anderson R. E., *Multivariate Data Analysis(7thed*), USA: Person Prentice Hall, 2010.