Advances in Social Science, Education and Humanities Research, volume 550
Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMED 2020)

Using the Rasch's Partial Credit Model to Analyze the Quality of an Essay Math Test

Sapti Wahyuningsih[1,*]

[1]*Universitas Negeri Malang, Indonesia*
Corresponding author. Email: sapti.wahyuningsih.fmipa@um.ac.id

ABSTRACT
The instrument in the form of an essay math test is one of the tools to measure the progress of the learning process or learning outcomes. The essay math test needs to be applied in education because students are trained to be independent, creative, innovative, and improve literacy skills. A good measurement model facilitates as much information as possible, for example, qualitative information, the accuracy of the interpretation of the instrument according to its function, provides a linear measure, can overcome missing data, and find data that is incorrect (misfits) or uncommon (outliers). Rasch's Partial Credit Model can identify the quality of the essay math test more accurately. The reliability value in Rasch modeling is indicated by the value of individual separation (person separation) and item separation. Output summary statistics to get information on the person and item reliability and Cronbach alpha. Analysis with the Rasch model produces a statistical analysis of suitability (fit statistics) which provides information that the data obtained ideally illustrates that people who have a high ability provide patterns of answers to items according to their level of difficulty. To analyze the quality of an essay math test using Rasch's Partial Credit Model, ministep software can be used.

Keywords: *Rasch's Partial Credit Model, reliability, an essay math test, ministep software.*

1. INTRODUCTION

The teaching and learning process in schools always involves educational assessment as a very important thing to do. Educational assessment is broader in scope than the test, which is more focused. The test is an evaluation procedure carried out by a teacher of the knowledge and skills of students to find out their performance using certain instruments. An essay math test is a test designed in the form of an essay to measure the basic mathematical abilities that students must master in certain subjects. The results of the math essay test need to be analyzed to determine the strengths and weaknesses of students in the mastery of mathematics so that more precise instruments can be prepared.

The approach in question is the application of Rasch model measurement to the raw data of the test results, the main objective of which is to produce a measurement scale with the same intervals that can provide accurate information about the test taker and the quality of the questions being tested. Referring to the researchers also using Rasch Modeling for instrument validation, see [1], [2], and [3].

Many researchers use the Rasch model to check the quality of the instrument, for example, to observe validity and reliability of instrument development can be seen in [4], [5], [6], [7], [2], and [8]. Not only checking validity and reliability, but the Rasch model can also be used to observe the abilities of students, for example, it can be seen in [9], [10], and [11]. Other research on Rasch models is Bloom's Separation [12], Rasch model analysis of negative symptom trajectories [13]. Development of instruments for attitude scale [14], interpreting and visualizing the unit of measurement [15], analysis of the psychometric [16], the effect of gender on teaching [17], substance problem scale [18], development of rating scale [19] and [20]. Other studies use Rasch measurement models to effect technology in learning [21], [22], and [23]. Some researchers in several countries use the Rasch model, for example in China [24], in South Africa [25], Indonesia [26], Singapore [27] and Malaysia [28].

The form of examination or test most commonly used by teachers is the written test. However, other forms of testing can also be used, such as oral or practical tests. A test must be valid, meaning that the test measures something that must be measured.

Although this concept seems simple, teachers usually forget about it. For example, the exam questions are arranged at the end of the collection of questions for a limited time. As a result, the subjects that are given in a complete and in-depth manner at the beginning of the lesson are, for example, not accommodated or even missed as questions on exams whose contents tend to contain only the final part of the subject matter. In other cases, if the desired learning outcome includes changes in knowledge, skills, and attitudes, then the questions made must also cover these three things.

There are various types of math tests, such as questions with a choice of answers, questions with correct or wrong answer choices, and essay questions. The type of description given to students has the same score pattern (the maximum number of scores is the same for each question) or the score pattern is not the same (the maximum number of scores is different for each question). In Rasch modeling to analyze data with different score patterns using the Partial Credit Model (PCM) Rasch measurement model. Research on PCM can be seen in [19] and [20].

This article will examine the use of the Partial Credit Model (PCM) Rasch measurement model to analyze an essay math test with a different maximum score for the graph theory course. Ministeps software tools are used to get output summary statistics, Write maps, misfit orders, measure order items, ICC charts, person measure orders, DIF plots, test information functions, and Partial credit scales to analyze the quality of an essay math test.

2. METHOD

The steps for using the Partial Credit Model (PCM) for Rasch modeling are described as follows.

- Preparation of raw data obtained from the results of an essay math test with different maximum scores. Code item utilizing questions with the same maximum score are given the same code and given different codes for different maximum scores. The results of the total value of each student (N = 54) from 5 items are stored in an excel form.
- Processing raw data into one column stored in .prn form
- Analyzing data on the ministep software application.
- Interpretation of the results of data processing.

3. RESULTS AND DISCUSSION

For example, the essay math test in the graph theory course are given to 54 students consisting of five essay items/problems. Problems no. 1 and no. 2 have the same maximum score, namely 6, problems no. 3 and 4 have the same maximum score namly 8 and problem no. 5 has a maximum score namely 9. In preparing the data to be processed with the PCM Rasch modeling, the problems 1 and 2 are given code A, the problems no. 3 and 4 are given code B and the problems no. 5 is coded C.

The output from the Ministep program application can be in the form of summary statistics, in Table 1. Summary statistics can be obtained by person reliability, item reliability, and Cronbach's alpha.

Table 1. Summary Statistics

	Total score	Measure	Infit	
			MNSQ	ZSTD
Mean	30.5	1.24	0.97	- 0.17
P.SD	3.5	2.26	0.86	1.21
S.SD	3.5	2.29	0.86	1.22

REAL	RMSE	0.97	TRUE SD	2.05	Separation	2.12	Person Reability	0.82
MODEL	RMSE	0.85	TRUE SD	2.10	Separation	2.47	Person Reability	0.86
Cronbach Alpha (Kr-20) Person Raw Score "Test" Reliability = 0.87								
REAL	RMSE	0.27	TRUE SD	0.62	Separation	2.29	Item Reability	0.84
MODEL	RMSE	0.26	TRUE SD	0.62	Separation	2.43	Item Reability	0.86

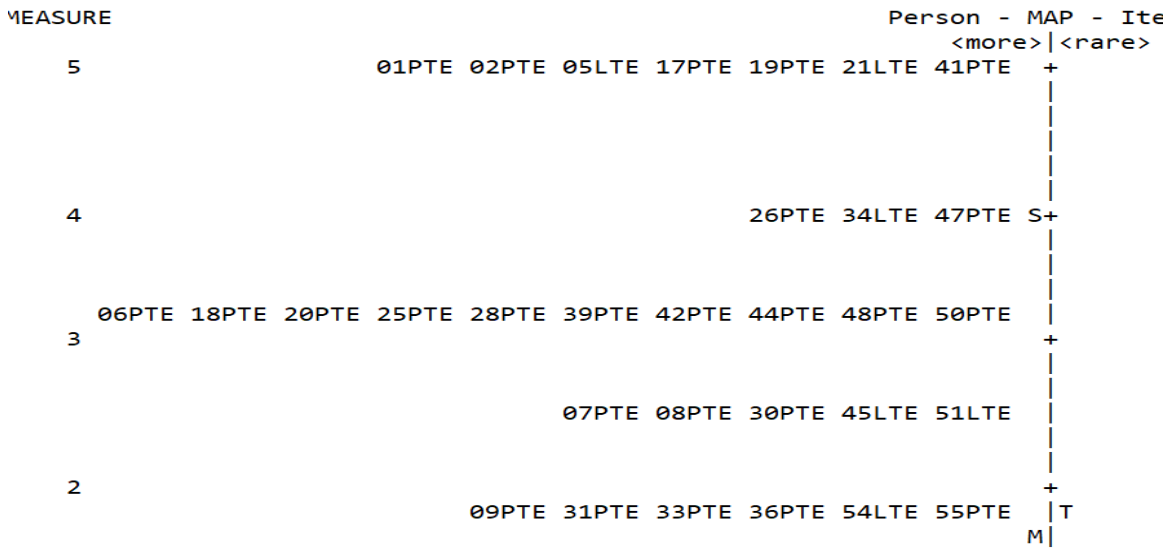
Table 2. Rating Scale Instrument Quality Criteria

Criterion	Poor	Fair	Good	Very Good	Excellent
Person Measurement Reliability	< 0.67	0.67 – 0.80	0.81 – 0.90	0.91 – 0.94	>0.94
Item Measurement Reliability	< 0.67	0.67 – 0.80	0.81 – 0.90	– 0.94	> 0.94
Fisher, W.P. Jr. (2007)					

Criteria for Cronbach Alpha:
<0.5: very poor, 0.5 - 0.6: poor, 0.6 - 0.7: fair, 0.7 - 0.8: good, > 0.8: very good

In Table 1, the statistical summary can be read that person reliability = 0.82 and item reliability = 0.84 based on the instrument quality criteria shown in Table 2, the criteria for person reliability and item reliability are both good. This result can be interpreted that the consistency of student answers and the quality of the question items in the instrument's reliability aspect is good. While Cronbach Alpha = 0.87 based on the criteria is very good. This means that the interaction between the person (student) and the item items as a whole is very good.

The advantages of Rasch modeling can be observing a map that describes the distribution of respondents' abilities and the distribution of difficulty levels of items with the same scale. This map in the Rasch model is depicted in the Write map. In the case of the Partial Credit Model the Write map produced is like the Rasch model in general. The write map output is shown in Figure 1, this result can be explained that the left shows the student's ability level and the right side shows the difficulty level of the items. In Figure 1. The Write map shows student identity number 01, female, graph theory class (01PTE), 02PTE,…, and 41PTE have the highest ability, while student identity number 49, female, graph theory class (49PTE) has the ability the lowest. On the Write map, it can be observed that item 3 (S3) has the highest difficulty level while item 2 (S2) has the lowest difficulty level.



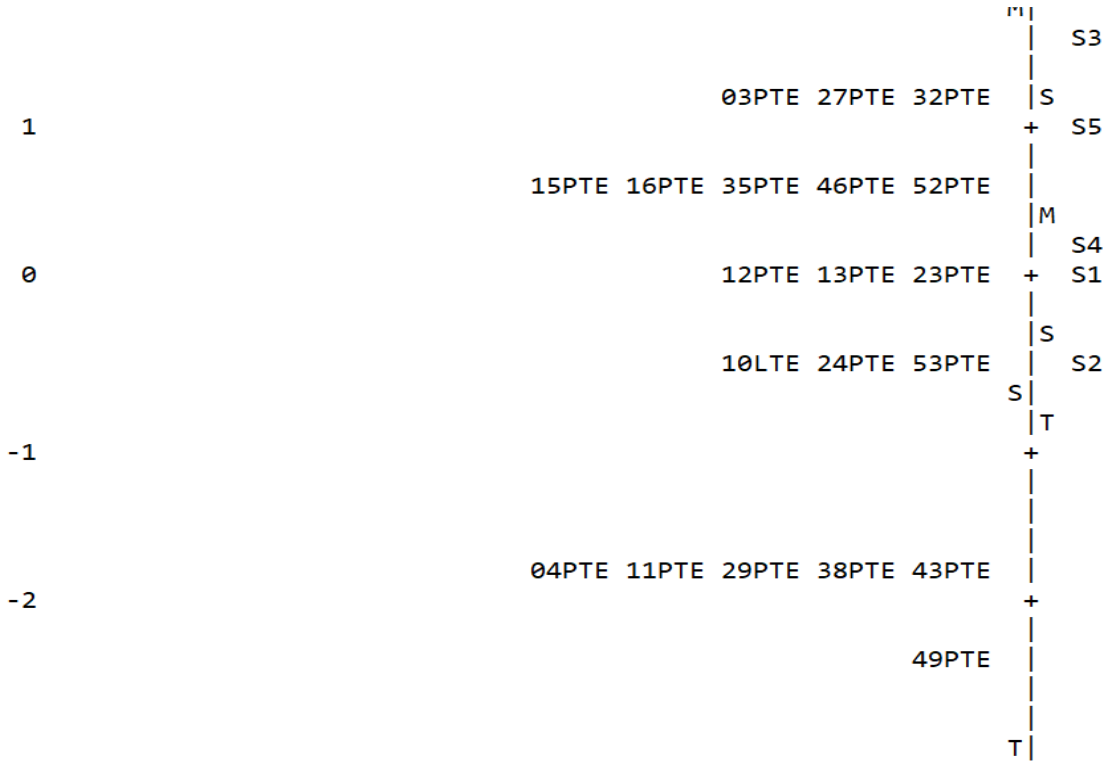


Figure 1. Write map

To see the level of suitability of the items with the ideal model, it can be seen in Table 3 the misfit order. Referring to [4-8] the criteria used to check the suitability of the items were 0.5 <Outfit MNSQ <1.5, -2.0 <Outfit ZSTD <+2.0 and 0.4 <Pt Measure Corr <0.85. The condition fulfills the criteria meaning that all the items given are understood by the student and no one has a misconception. It can be seen that in Table 3 the misfit order, the three criteria are accepted, namely that there are no outliers or misfit items. This is also supported by the graph of the expected score Item Characteristic Curve (ICC) which can be seen in Figure 2 all responses (marked with an x) are located in the infit confidence space curve and the outfit follows the ideal model line curve.

In Table 3, the misfit order can be seen that the point measure correlation (PT-Measure Corr) is all positive, this indicates that the item has construct validity. Discussion of validity can be seen in [4] and [7-8].

Table 3. Misfit Order

Total Count	PT-Measure Corr	Outfit	
		MNSQ	ZSTD
54	0.77	1.27	1.28
54	0.72	1.19	0.99
54	0.82	0.97	- 0.05
54	0.86	0.84	- 0.73
54	0.87	0.68	- 1.91

The level of difficulty of the items can be seen in Table 4. Measure order. The logit value can be seen in the measure column which is sorted from the highest to the lowest logit value. The highest logit value indicates the highest level of problem difficulty. It can be seen that item S3 with item code B has the highest difficulty level while S2 has the lowest logit value indicating the easiest item.

Table 4. Item Measure Order

Total score	Total count	Measure	Item	G
331	54	0.97	S3	B
415	54	0.54	S5	C
350	54	- 0.18	S4	B
267	54	- 0.41	S1	A
274	54	- 0.92	S2	A
Mean 327.4	54	0.00		
P.SD 54.2	0.0	0.67		

To see the suitability of the model in addition to the results shown in Table 3 for the misfit order, the ICC expected score graph can be used. Figure 2 graph ICC problem no 1 (S1), the red curve is the ideal model line curve, the outfit confidence space curve is located on the right and the infit confidence space curve is the left side of the ideal model line curve. If all responses marked (*x*) lie around the curve of the ideal model line, nothing outside the infit confidence space curve and the outfit confidence space curve means the model is accepted.

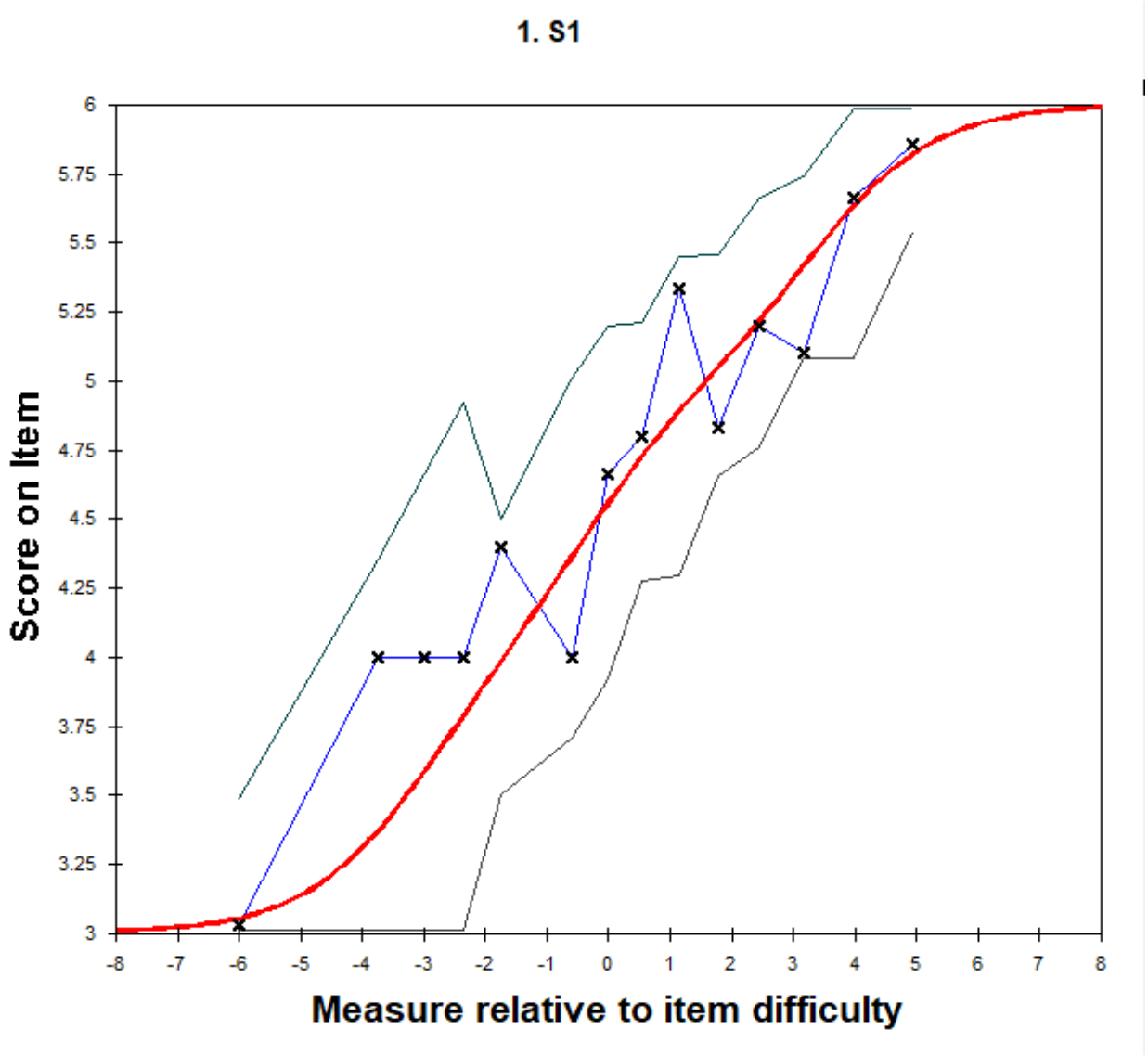


Figure 2. Graph ICC Problem no 1

An item needs to be checked whether there are items that contain bias for certain groups, for example, gender, domicile, class, or major. The output selected in the ministep application is a DIF item. Writing DIF = $ S3W1 means that the grouping of respondents is based on column 3, which is based on the gender of male (L) and female (P). Table 5. The results of the DIF item output and Figure 3 DIF plots are shown with different color curves.

Table 5 DIF group gender

Person/Class	DIF Measure	DIF S.E	Prob.
L	- 0.96	0.81	0.1941
L	- 0.96	0.81	0.6615
L	0.47	0.71	0.7151
L	0.47	0.71	0.1329
L	0.84	0.69	0.7207
P	- 0.34	0.28	0.1941
P	- 0.92	0.29	0.6615
P	1.04	0.26	0.7151
P	- 0.28	0.27	0.1329
P	0.50	0.27	0.7207

"L : male, P: female"

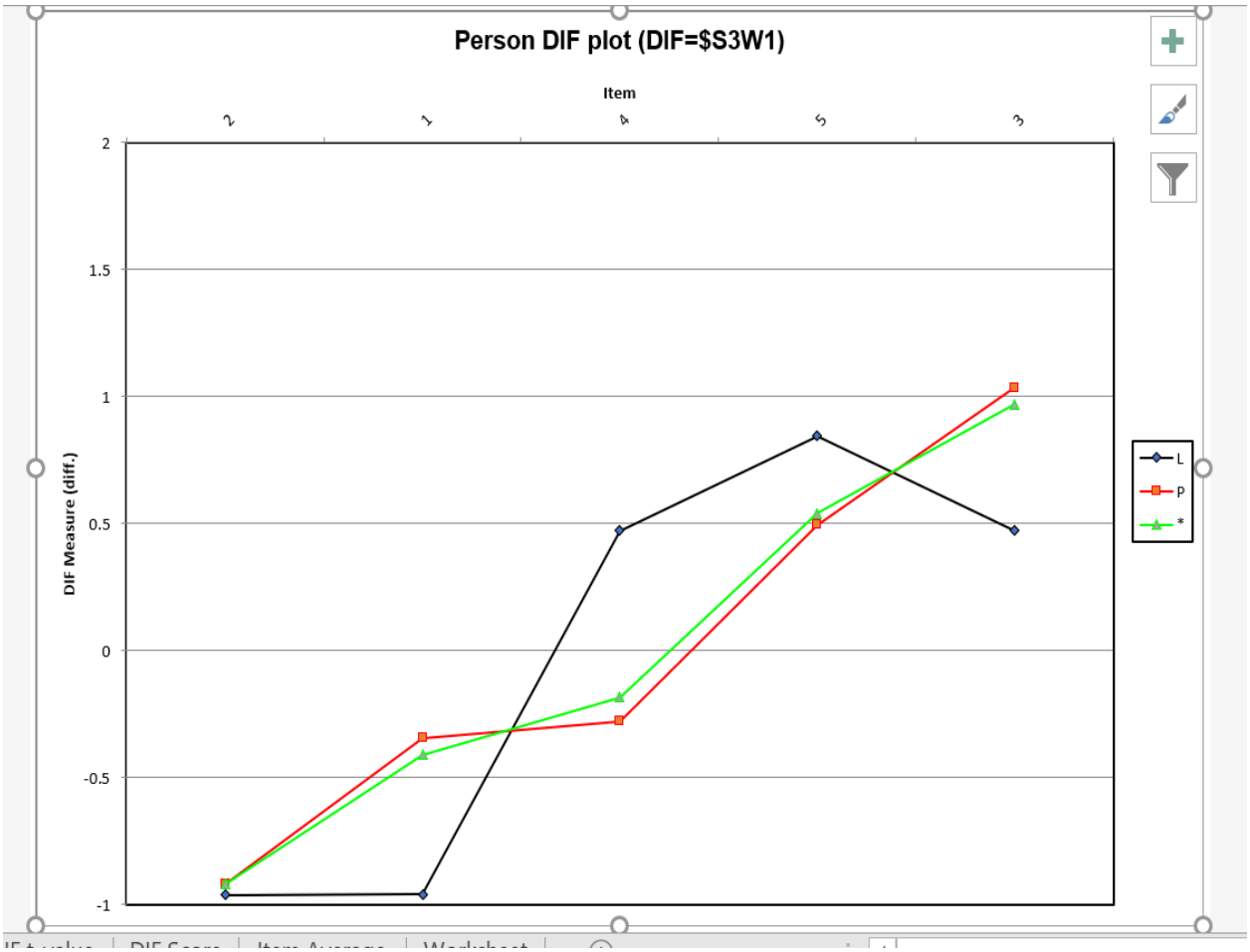


Figure 3. Plot DIF Gender

The use of partial credit models for rating scale development can be seen in [19]. Figure 4 shows the probability curve for the PCM category. Probability curves are useful for checking the level of difficulty of items that may not be suitable for respondents.

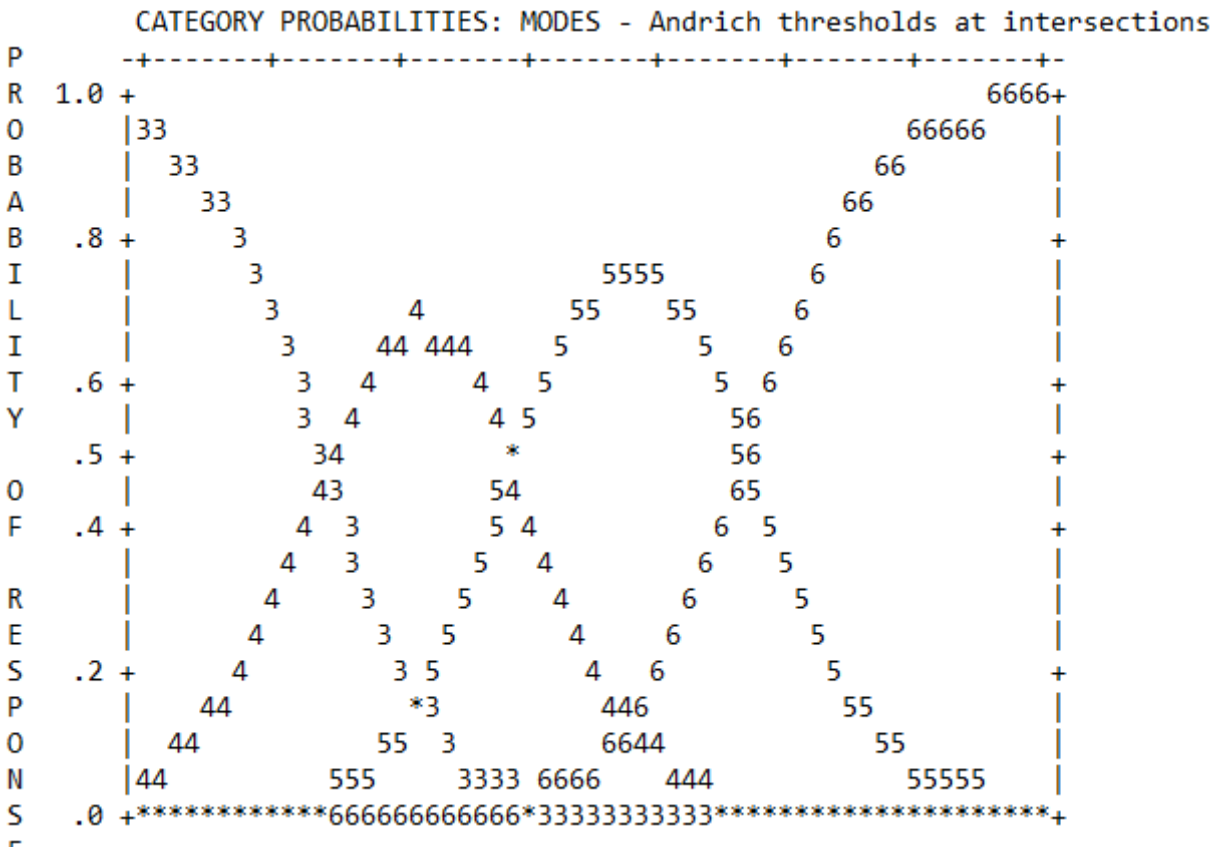


Figure 4. The probability curve for the PCM category

4. CONCLUSION

The Rasch Modeling Partial Credit Model (PCM) can be used to analyze the quality of the essay math test. The Cronbach alpha, person, and item reliability values can be interpreted from the output summary statistics. The distribution of student abilities and the level of difficulty of the items can be interpreted from the output of the Write map. These results can be used for grouping the level of student ability which is distributed linearly with the difficulty level of the items. The use of Rasch modeling in addition to educational assessments can also be used to analyze the quality of research instruments.

ACKNOWLEDGMENTS

This article is a part of research funded by PNBP UM, contract no: 4.3.475/UN32.14.1/LT/2020, Thank you for Universitas Negeri Malang for Funding the Research.

REFERENCES

[1] R. Mohd. Yasin, F. A. N. Yunus, R. C. Rus, A. Ahmad, and M. B. Rahim, 'Validity and Reliability Learning Transfer Item Using Rasch Measurement Model', *Procedia - Social and Behavioral Sciences*, vol. 204, pp. 212–217, Aug. 2015, DOI: 10.1016/j.sbspro.2015.08.143.

[2] G. H. Alnahdi, 'Rasch validation of the Arabic version of the Teacher Efficacy for Inclusive Practices (TEIP) scale', *Studies in Educational Evaluation*, vol. 62, pp. 104–110, Sep. 2019, DOI: 10.1016/j.stueduc.2019.05.004.

[3] H. Garcia Claro *et al.*, 'Rasch model of the GAIN substance problem scale among inpatient and outpatient clients in the city of São Paulo, Brazil', *Addictive Behaviors Reports*, vol. 2, pp. 55–60, Dec. 2015, DOI: 10.1016/j.abrep.2015.08.001.

[4] E. R. Jacob, C. Duffield, and A. M. Jacob, 'Validation of data using RASCH analysis in a tool measuring changes in critical thinking in nursing students', *Nurse Education Today*, vol. 76, pp. 196–199, May 2019, DOI: 10.1016/j.nedt.2019.02.012.

[5] N. Wongpakaran, T. Wongpakaran, and P. Kuntawong, 'A short screening tool for borderline personality disorder (Short-Bord): Validated by Rasch analysis', *Asian Journal of Psychiatry*, vol. 44, pp. 195–199, Aug. 2019, DOI: 10.1016/j.ajp.2019.08.004.

[6] J. M. Pérez-Mármol and T. Brown, 'An Examination of the Structural Validity of the Maslach Burnout Inventory-Student Survey (MBI-SS) Using the Rasch Measurement Model', *Health Professions Education*, vol. 5, no. 3, pp. 259–274, Sep. 2019, DOI: 10.1016/j.hpe.2018.05.004.

[7] D. Marengo *et al.*, 'Examining the validity of the multiple-sclerosis walking scale-12 with Rasch analysis: Results from an Italian study', *Multiple Sclerosis and Related Disorders*, vol. 36, p. 101400, Nov. 2019, DOI: 10.1016/j.msard.2019.101400.

[8] R. Mohd. Yasin, F. A. N. Yunus, R. C. Rus, A. Ahmad, and M. B. Rahim, 'Validity and Reliability Learning Transfer Item Using Rasch Measurement Model', *Procedia - Social and Behavioral Sciences*, vol. 204, pp. 212–217, Aug. 2015, DOI: 10.1016/j.sbspro.2015.08.143.

[9] M. Wati, S. Mahtari, S. Hartini, and H. Amelia, 'A Rasch Model Analysis on Junior High School Students' Scientific Reasoning Ability', *Int. J. Interact. Mob. Technol.*, vol. 13, no. 07, p. 141, Jul. 2019, DOI: 10.3991/ijim.v13i07.10760.

[10] S. W. Chan, Z. Ismail, and B. Sumintono, 'A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics', *Procedia - Social and Behavioral Sciences*, vol. 129, pp. 133–139, May 2014, DOI: 10.1016/j.sbspro.2014.03.658.

[11] M. N. Mamat, P. Maidin, and F. Mokhtar, 'Simplified Reliable Procedure for Producing Accurate Student's Ability Grade Using Rasch Model', *Procedia - Social and Behavioral Sciences*, vol. 112, pp. 1077–1082, Feb. 2014, DOI: 10.1016/j.sbspro.2014.01.1272.

[12] I. Asshaari, H. Othman, H. Bahaludin, N. A. Ismail, and Z. M. Nopiah, 'Appraisal on Bloom's Separation in Final Examination Question of Engineering Mathematics Courses using Rasch Measurement Model', *Procedia - Social and Behavioral Sciences*, vol. 60, pp. 172–178, Oct. 2012, DOI: 10.1016/j.sbspro.2012.09.364.

[13] L. Baandrup *et al.*, 'Rasch analysis of the PANSS negative subscale and exploration of negative symptom trajectories in first-episode schizophrenia – data from the OPTiMiSE trial', *Psychiatry Research*, vol. 289, p. 112970, Jul. 2020, DOI: 10.1016/j.psychres.2020.112970.

[14] B. Boroel, V. Aramburo, and M. Gonzalez, 'Development of a Scale to Measure Attitudes Toward Professional Values: An Analysis of Dimensionality Using Rasch Measurement', *Procedia - Social and Behavioral Sciences*, vol. 237, pp. 292–298, Feb. 2017, DOI: 10.1016/j.sbspro.2017.02.079.

[15] D. C. Briggs, 'Interpreting and visualizing the unit of measurement in the Rasch Model', *Measurement*, vol. 146, pp. 961–971, Nov. 2019, DOI: 10.1016/j.measurement.2019.07.035.

[16] C. S. da Conceição, M. G. Neto, A. C. Neto, S. M. D. Mendes, A. F. Baptista, and K. N. Sá, 'Analysis of the psychometric properties of the American Orthopaedic Foot and Ankle Society Score (AOFAS) in rheumatoid arthritis patients: application of the Rasch model', *Revista Brasileira de Reumatologia (English Edition)*, vol. 56, no. 1, pp. 8–13, Jan. 2016, DOI: 10.1016/j.rbre.2014.12.003.

[17] J. F. Ehrich, S. Woodcock, and C. West, 'The effect of gender on teaching dispositions: A Rasch measurement approach', *International Journal of Educational Research*, vol. 99, p. 101510, 2020, DOI: 10.1016/j.ijer.2019.101510.

[18] H. Garcia Claro *et al.*, 'Rasch model of the GAIN substance problem scale among inpatient and outpatient clients in the city of São Paulo, Brazil', *Addictive Behaviors Reports*, vol. 2, pp. 55–60, Dec. 2015, DOI: 10.1016/j.abrep.2015.08.001.

[19] C. Van Zile-Tamsen, 'Using Rasch Analysis to Inform Rating Scale Development', *Res High Educ*, vol. 58, no. 8, pp. 922–933, Dec. 2017, DOI: 10.1007/s11162-017-9448-0.

[20] M. H. Sandham, O. N. Medvedev, E. Hedgecock, I. J. Higginson, and R. J. Siegert, 'A Rasch Analysis of the Integrated Palliative Care Outcome Scale', *Journal of Pain and Symptom Management*, vol. 57, no. 2, pp. 290–296, Feb. 2019, DOI: 10.1016/j.jpainsymman.2018.11.019.

[21] N. F. Hassan, S. Puteh, A. M. Sanusi, and N. H. C. M. Zahid, 'Student Perspective on Technology-Enabled/Enhanced Active Learning in Educational: Rasch Measurement Model', *Int. J. Onl. Eng.*, vol. 16, no. 06, p. 34, May 2020, DOI: 10.3991/ijoe.v16i06.13575.

[22] I. Ismail, R. Mohammed Idrus, and S. S. Mohd Johari, 'Acceptance on Mobile Learning via SMS: A Rasch Model Analysis', *Int. J. Interact. Mob. Technol.*, vol. 4, no. 2, pp. 10–16, Apr. 2010, DOI: 10.3991/ijim.v4i2.1144.

[23] A. Salman and A. Abd. Aziz, 'Evaluating user Readiness towards Digital Society: A Rasch Measurement Model Analysis', *Procedia Computer Science*, vol. 65, pp. 1154–1159, 2015, DOI: 10.1016/j.procs.2015.09.028.

[24] C. U. Krägeloh, G. Y. Wang, Q. Zhao, O. N. Medvedev, Y. Wu, and M. A. Henning, 'Revised Competitiveness Index for use in China: Translation and Rasch analysis', *International Journal of Educational Research*, vol. 90, pp. 78–86, 2018, DOI: 10.1016/j.ijer.2018.05.008.

[25] M. Makhubela, 'Using the Trauma Symptom Checklist for Children-Short form (TSCC-SF) on abused children in South Africa: Confirmatory factor analysis and Rasch models', *Child Abuse & Neglect*, vol. 98, p. 104241, Dec. 2019, DOI: 10.1016/j.chiabu.2019.104241.

[26] B. Setiawan, M. Panduwangi, and B. Sumintono, 'A Rasch analysis of the community's preference for different attributes of Islamic banks in Indonesia', *Int J of Social Economics*, vol. 45, no. 12, pp. 1647–1662, Dec. 2018, DOI: 10.1108/IJSE-07-2017-0294.

[27] N. Thanh and D. N. F. Seong, 'Applying the Rasch Model to Investigate Singapore Principals' Instructional Leadership Practices', p. 26.

[28] N. Zahir and B. Sumintono, 'Perceptions on Influence Tactics among Leaders in the Ministry of Education Malaysia: An Application of The Many Facets Rasch Model', p. 13.