The Quality of Non-Routine Math Questions for Junior High School in Partial Credit Model Scaling

Sugeng
Mathematic Education
Department, Faculty of Theacher
Training and Education,
Mulawarman University
Samarinda, Indonesia
sugeng@fkip.unmul.ac.id

Aulia Ariski Asmawati
Mathematic Education
Department, Faculty of Theacher
Training and Education,
Mulawarman University
Samarinda, Indonesia
aulia_arisky@yahoo.com

Yulia Dewi Arief Fanti
Mathematic Education
Department, Faculty of Theacher
Training and Education,
Mulawarman University
Samarinda, Indonesia
yuliadewiaf.ydaf@gmail.com

Abstract --- **This study aims to obtain a mathematics learning outcome measurement tool that meets the quality requirements of the questions at the junior high school level and is developed using a modern test theory approach (Item Response Theory, IRT). This study involved samples of students in grades VII, VIII, and IX at SMP 2 and SMP 5 Samarinda in 2017. Determination of the research sample using simple random sampling technique at the school level. PCM scaling uses non-routine items with *polytomous* scoring. Data analysis techniques to find out the quality of the items using two ways, namely quantitative and qualitative methods. The results showed that mathematical questions with PCM scaling met the expected quality. The results showed that from 20 questions there were 13 items in the quality category**.

Keywords: ***Quality of questions, Non-routine math questions, IRT, PCM***

I. INTRODUCTION

Mathematics is one of the mandatory lessons for junior high school students. Mathematical material is arranged hierarchically. For the same branch of mathematics, the material in the upper class is a continuation of the material in the class below. In addition, the competencies achieved by students are determined based on the mastery and skills of students in teaching materials according to the grade level. These competencies can be revealed through non-routine mathematical questions. Ministry of Education and Culture **[2]** revealed that knowledge competence is not only to conceptual understanding but to application through procedural knowledge in solving mathematical problems. Therefore, to get accurate information about students' mathematical competencies, a quality measuring instrument is needed.

Various measuring instruments in measurement activities are in the form of tests and non-tests **[3] [4]**. In the practice of evaluation, there are three interrelated aspects, namely between test, measurement, and evaluation **[5]**. In principle, the measurement aims to determine the characteristics of an object related to cognitive, effective, and psychomotor aspects **[4] [6]**. The measurement activity is one of the main activities in evaluating student learning outcomes, especially mathematics learning. To get the right information about the characteristics of an object, a good measuring instrument is needed, namely a measuring instrument that meets the requirements.

Measurement as a systematic activity to obtain information in quantitative form (numbers). Determination of numbers in measurements as an attempt to describe the characteristics of an object **[7] [8]**. The measurement results must have the smallest possible error **[4]** .. The source of the measurement error lies in the measuring instrument, how to measure, the person making the measurement, the person who measured his psychological attributes, and the condition or environment when the measurement was made. Kusaeri & Suprananto **[9]** revealed that measurement has three characteristics; namely (1) measurement as a comparison between attributes measured by measuring instruments; (2) the results of measurements are quantitative; (3) the measurement results are descriptive.

Item Response Theory (IRT) as a measurement approach based on item responses to determine the latent characteristics of an object **[10]**. IRT uses unidimensional assumptions and local independence assumptions. The assumed dimensions pertain to the accuracy of a measuring instrument that only measures one type of ability, or the substance to be measured must be one dimension. In addition, the function in θ is used in IRT models for dichotomous items, both 1-PL, 2-PL, and 3-PL forms **[11]**, and also to determine the probability price of the respondent's ability to answer correctly, the amount of information item Ii (θ), and test information I (θ) **[12]**. Therefore, the IRT seeks to develop an analysis that results in estimation of a person's ability without being influenced by its measurement tools.

Partial Credit Model *(PCM)* is one of the IRT models developed by Masters **[1]** based on dichotomous responses (two alternative answers; Rasch models) to more than two alternative answers. The algebraic model of PCM **[13]** as follows:

$$P_{nix} = \frac{1}{1+\sum_{k=1}^{m_i}\exp\sum_{j=1}^{k}(\beta_n - \delta_{ij})}, \text{ for } x = 0 \qquad (1)$$

$$P_{nix} = \frac{\exp\sum_{j=1}^{x}(\beta_n - \delta_{ij})}{1+\sum_{k=1}^{m_i}\exp\sum_{j=1}^{k}(\beta_n - \delta_{ij})} \qquad (2)$$

for $x = 1, 2, 3, \ldots, m_i$

PCM scaling in mathematical material is gradation, starting at 0, 1 and 2 for each item. Determination of the scale according to the weight of the response, which starts from 2 decreases according to the concepts in the response of each alternative choice answer on the item. Zero weight (0) is given to respondents who did not provide answers to the items.

Mathematics is an exact branch of science, which is systematically organized mainly with regard to numbers and calculations, and has characteristics: (a) abstract study objects, (b) rely on agreements, (c) deductive thinking patterns, (d) have empty symbols of meaning, (e) paying attention to the universe of speech, and (f) consistent in the system **[14]**. Mathematics curriculum that has been recommended establishes that thinking (thinking ability) and problem solving (problem solving) is associated with all topics of Mathematics (numbers & operations, statistics, measurement, probability, geometry, algebra) and its contents (content) associated with life (living) , working, and solving problems **[15]**. With regard to the need for measurement, Mathematical questions are divided into two, namely (a) routine questions, and (b) non-routine questions **[16]**. For the completion of non-routine questions, one of the completion techniques from Polya **[17]**, namely using sequential stages: (1) understanding the problem, (2) devising plan, (3) carrying out the plan, and (4) looking back.

The quality of the math test instrument refers to the accuracy of the measurement results using the questioned instrument. The level of difficulty refers to the proficiency of examinees 50% (or smaller, depending on the IRT model applied) which is expected to be able to answer the questions correctly. According to **[18]**, if the *ability* value of a group is transformed so that the mean is 0 and the standard deviation is 1, then the value of *bi* changes from –2.0 to +2.0. The b_i value is close to -2.0 correspond to items that are very easy; while the *bi* value approaching 2.0 corresponds to items that are very difficult.

The higher the differentiating power of a question shows that the differentiation of items between test participants' construct rates is increasingly different. Theoretically, the item discrimination parameter is defined as scala (–∞, ∞). The grain discrimination index marked negative is discarded **[18]**. Usually the range for grain discrimination parameters is (0, 2). The higher the price of a_i in a function of the grain characteristics is the more "steep", and the lower the price of a_i in the function of the grain characteristics is shown to gradually decrease as the function in *ability.*

II. METHODS

This research was conducted in July - October 2017 in Samarinda at the junior high school level. for all grade levels. The study sample involved 82 students of class VII, 71 students of class VIII, and 59 students of class IX of SMP 2 and SMPN 5 Samarinda. Determination of stratified random sampling. With regard to the *Principal Component Analysis (PCA)* type Factor Analysis technique in investigating the construct validity.

Data collection techniques using tests. But the test questions were compiled by the researchers using *Partial Credit Model (PCM)* scaling, based on the question grid with the grade VII Mathematics in grade VIII, class VIII, and class IX each of 20 items. The grid for the preparation of test items follows the pattern of **[2]**. The *PCM* type Mathematics test instruments involve Numbers, Algebra, and Geometry & Measurement material. In accordance with the nature of the *PCM* model, scoring of answers uses category scores, namely 0, 1, and 2.

Data analysis in this study uses the QUEST program from Adam & Kho **[19]**. Determination of fit items analysis results with QUEST, as a whole based on the INFIT Mean of Square (INFIT MNSQ) value along with the standard deviation; or INFIT Mean of INFIT value t **[19]**.

III. RESULTS AND DISCUSSION

1. Partial Credit Model of Items

The *PCM* items are compiled using all SMP material, starting from the material grades VII to class IX .. This combination is used to find out the extent to which the questions are with the basic concepts of mathematics that have been studied (for students in higher grade levels), or those being studied (for students at lower grade levels), they are able to do it.

2.Construct Validity of Items

The testing of construct validity **[20]** aims to show that this measure tends to measure one dimension (dimensions) based on the data collected through measurement with *PCM*. Testing of construct validity is done using factor analysis. The main objective is to prove that as a whole the items of this instrument are likely to lead only to one dimension. The analysis was carried out using the *SPSS* program assistance. The test value is indicated by the Coefficient of *KMO (Kaiser-Meyer-Olkin Measure) and Bartlett's test*. The results of testing data for class VII, class VIII, and class IX with a full factor analysis are shown in Table 1 below.

Table 1. *KMO* Coefficient of Data

Grades	Koefisien *KMO* *and Bartlett's test*	*Sig.*
VII	0.720	0.000
VIII	0.942	0.000
IX	0.556	0.000

Based on Table 1, the coefficients of *KMO and Bartlett's test* for data class VII, class VIII, and class IX, respectively are 0.720, 0.942, and 0.556, and each with significance Sig. = 0,000. The price coefficient has exceeded 0.50 with a significance far below 0.05 (p <α). This shows that the collection of variables (the results of the three math tests) can be further processed. The results of factor analysis using the *PCA* method are graphically reinforced by the results of the Scree Plot. As a visualization, Scree Plot is presented for the following class VIII data.

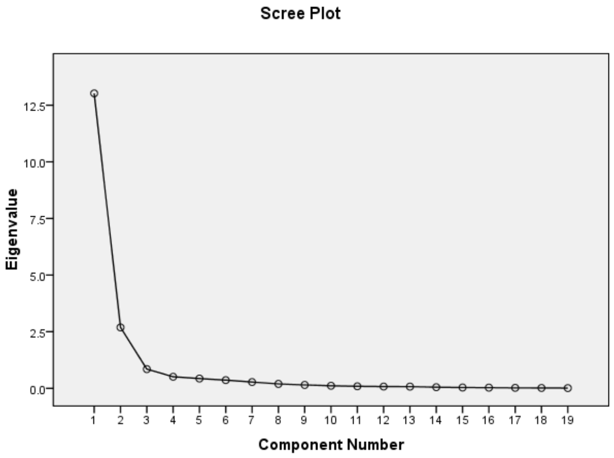

Figure 1. Scree Plot Results of Factor Analysis of Grade VIII Mathematics Learning Outcomes

The results of factor analysis showed that the overall PCM Mathematics items of class VIII students (19 items) grouped into 2 factors. In the *Scree Plot* and the Table of *Total Variance Explained* in class VIII, the eigenvalue weight of each component of the statement item is different from each other, at most 13,024 then decreases sharply to 2,690, and so on the smaller the weight.

Of the two factors that have the largest eigenvalue weight, there is one dominant factor, which is indicated by the longest line *(*on the Scree Plot) with an eigenvalue weighting of 13,024. This condition shows the existence of one dominant component *(factor)* in measuring *PCM* mathematics learning outcomes; this means that this measuring instrument tends to measure one factor. Thus, measuring the learning outcomes of mathematics class VIII with *PCM* tends to measure only one dimension. Thus construct validity is fulfilled

3.Difficulty Index

Based on the results of the analysis with the QUEST program, the difficulty index for each test item was obtained as follows.

Table 2. The Difficulty Index of Items

```
 ------------------------------------
   Item Estimates (Difficulty and
   Taus)
   in input Order all on all (N = 212
   L = 20 Probability Level= .50)
  -----------------------------------
  ITEM NAME | DIFFCLTY  TAU/S      |
-------------------------------------
 1   item 1  |   .43    -.19    .19 |
 2   item 2  |   .54   -1.13   1.13 |
 3   item 3  |  -.19     .05   -.05 |
 4   item 4  |   .69     .08   -.08 |
 5   item 5  |   .41    -.01    .01 |
 6   item 6  |   .61    -.26    .26 |
 7   item 7  |  1.02    -.88    .88 |
8   item 8  |  2.53    -.86    .86 |
9   item 9  |   .60    -.56    .56 |
10  item 10 |   .89    -.34    .34 |
11  item 11 |  -.94    -.26    .26 |
12  item 12 |  -.94     .20   -.20 |
13  item 13 | -1.23     .15   -.15 |
14  item 14 |  -.66    -.56    .56 |
15  item 15 | -1.10     .17   -.17 |
16  item 16 |  -.42    -.39    .39 |
17  item 17 |  -.74    -.35    .35 |
18  item 18 |  -.72    -.49    .49 |
19  item 19 |  -.24    -.85    .85 |
20  item 20 |  -.53    -.21    .21 |
------------------------------------
Mean                 |    .00      |
SD                   |    .93      |
====================================
```

Difficulty index has a range of -2 ≤ b ≤ 2. Based on these criteria, 19 items of mathematics are categorized as good, meaning that they fulfill those requirements, namely items number 1, 2, 3, 4, 5, 6, 7, 9, 10 , 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20. However, there is 1 item whose difficulty index does not fulfill, namely item 8 with a Difficulty index of 2.53.

4.Test Reliability

One of the results of the analysis with the QUEST program shows the reliability coefficient of the math test, as follows.

```
---------------------------------
Item Analysis Results for Observed
Responses all on all (N = 212 L = 20
Probability Level= .50)
------------------------------------
Mean test score        20.30
Standard deviation      6.46
Internal Consistency     .79
=================================
```

Based on the results of this analysis obtained the reliability coefficient (internal consistency) of 0.79. This coefficient is categorized quite high.

5.The Discrimination Index

The discrimination index was assessed through biserial-point values. The results of the analysis with IRT show that the item index discrimination index is spread in intervals 0 ≤ a ≤ 2. The QUEST analysis results show the following:

Table 3. The Discrimination Index of Item

Nomor Butir	Pt-Biserial	category
1.	-.12	Poor
2.	.48	Good
3.	.59	Good
4	.60	Good
5	.17	Good
6	-.05	Poor
7	.17	Poor
8	-.02	Poor
9	.57	Good
10	.52	Good
11.	.49	Good
12	.38	Good
13.	.46	Good
14.	.52	Good
15	.53	Good
16.	.65	Good
17.	.58	Good
18.	.46	Good
19.	.23	Mediocre
20.	.35	Good

Based on the table above, there are 3 items (items 1, 6, 7, 8) in the poor category, and 1 (item). mediocre category (item 19).

6. Fit-Item Analysis

Graphically (according to the ICC, Item Characteristic Curve), it is determined that an item or test or case or person is declared fit with the model, if it is in the MNSQ INFIT range from 0.77 to 1.30 (Adam & Kho, 1996). Based on the graphic above, the results show that (1) the items that were discarded were items 1, 6, 8, and 16. Thus, there were still 16 items categorized as items that were fit, meaning that they were suitable for the model used, namely PCM. Thus the items that are fit are items 2, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, and 20. (See Table 5).

Based on various results of analysis of the components of quality measuring instruments, it can be summarized in Table 4.

Table 4. Result of Items Analysis

Parameters	Result of items Analysis
1. Difficulty index	2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 (delete: 1, 8)
2. Construct Validity	The KMO and Bartlett's test coefficient requirements are met
3. Discrimination Index	2, 3, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 (delete:: 1, 6, 7, 8, 19)
4. Reliability	0,79
5. Fit-item	2, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20. (delete:: 1, 6, 8, 16)
Conclution	2, 3, 4, 5, 10, 11, 12, 13, 14, 15, 17, 18, 20

The results of the item analysis of the math test questions for the non-routines type questions showed that there were 13 items which were of good quality of the 20 items available. The items in the question fulfill the requirements as good items. There are five item requirements that are good in this study, namely the construct validity of the item, the difficulty level of the item, the distinguishing power, and the fit-item.

Table 5. Results of fit-item analysis

```
--------------------------------------------------------------------------
--
Item Fit                                            -          7/11/18
5:29
all on all (N = 212 L = 20 Probability Level= .50)
--------------------------------------------------------------------------
--
INFIT
MNSQ           .63        .71         .83       1.00       1.20       1.40
1.60
--------------+---------+---------+---------+---------+---------+---------
+-
  1 item 1                   .              |              .           *
  2 item 2                   .             *|              .
  3 item 3                   .              |*             .
  4 item 4                   .*             |              .
  5 item 5                   .              |         *    .
  6 item 6                   .              |              .        *
  7 item 7                   .              |    *         .
  8 item 8                   .              |              *
  9 item 9                   .       *      |              .
 10 item 10                  . *            |              .
 11 item 11                  .          *   |              .
 12 item 12                  .              |   *          .
 13 item 13                  .     *        |              .
 14 item 14                  .  *           |              .
 15 item 15                  .*             |              .
 16 item 16            *     .              |              .
 17 item 17                  . *            |              .
 18 item 18                  .    *         |              .
 19 item 19                  .             *|              .
 20 item 20                  .           *  |              .
==========================================================================
```

The distinguishing parameters that are determined using point-biserial correlation, are generally used in item analysis with the *(classical theory test approach)*. However, the point-biserial coefficient is the result of analysis with the QUEST program, which is one of the tools for testing with a modern res approach. Therefore, the differentiating requirements using biserial points are only supporting.

The need for test questions for school lessons, especially the making of question banks, this analysis technique is quite adequate. This means that this analysis activity can be developed to provide other lesson items besides math lessons. In the analysis activities with the modern test theory approach *(IRT)*, the quality of the items in the math test questions was only assessed using one application program, namely *QUEST*, as a result of the limitations of the available application programs. The quality analysis of the items of this test will result in better results if the application programs used are more diverse

IV. CONCLUSION

The quality of the test items for a lesson, especially Mathematics, is done using the Modern Test Theory (Item Response Theory) approach. The information used to state the quality of the test items is more complex than using classical test theory. This study can be used for making bank questions about the field of mathematics.

REFERENCES

[1] Masters, G. N. A Rasch model for partial credit scoring. *Psychometrica*, 1982, *47*(2), 149–174.

[2] Kemdikbud. *Matematika: Buku guru. Untuk SMP/MTs Kelas VII*. Jakarta. . 2016

[3] Sumadi Suryabrata. *Pengembangan alat ukur psikologis*. Yogyakarta: ANDI. 2000

[4] Djemari Mardapi. *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendikia Press. 2008

[5] Mehrens, W. A. & Lehmann, I. J. *Measurement and evaluation in education and psychological* (2nd ed.). New York: Holt, Rinehart, and Winston. 1978

[6] Djemari Mardapi. Analisis butir dengan teori tes klasik dan teori tes respons butir. *Laporan Penelitian*. Yogyakarta: IKIP Yogyakarta. 1994

[7] Nunnaly, J. C. *Psychometryc theory*. New York: McGraw-Hill. 1978

[8] Reynold, C. R., Livingstone, R. B., & Willson, V. *Measurement and assessment in education*. New Jersey: Pearson Education, Inc. 2010

[9] Kusaeri & Suprananto. *Pengukuran dan penilaian pendidikan*. Yogyakarta: Graha Ilmu. 2012

[10] Hulin, C. L., Drasgow, F., & Parsons, C. K. *Item response theory: Application to psychological theory*. Homewood, IL: Dow Jones-Irwin. 1983.

[11] Djemari Mardapi. *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendikia Press. 2008.

[12] Embretson, S. E. & Reise, S. P. *Item response theory for psychologist.* Mahwah, NJ: Lawrence Erlbaum Associates. 2000

[13] Masters, G. N. Partial credit model. Dalam G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 98–109). Amsterdam: Pergamon. 1999

[14] Soedjadi, R. *Kiat pendidikan matematika di Indonesia*. *Konstatasi keadaan masa kini menuju harapan masa depan.* Jakarta: Depdiknas, Ditjen Dikti. 2000

[15] Kennedy, L. M., Tipps, S, & Johnson, A. *Guiding childrens's learning of mathematics* (11th ed.). Belmon VA: Thomson Wardswoth. 2008

[16] Possamentier, A. S, & Stepelman, J. T*eaching secondary school mathemate matics: Teaching and enrichment units* (3nd.ed.). Columbus OH: Merril Publishing company. 1990

[17] Polya, G. *How to solve it. A new aspect of mathematical method*. Princeton, NJ: Princeton University Press. 1973

[18] Hambleton, R. K., Swaminathan, H., & Rogers. H. J. *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications. 1991

[19] Adams, R.J. & Kho, Seik-Tom. *Acer quest version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research. 1996

[20] **[20]** Caroll, J. B. Measurement of abilities constructs. Dalam *Construct Validity in Psychological Measurement* (pp. 23-41). Proceedings of a Colloqium on Theory and Application in Education and Employment. Princeton, NJ: ETS. 1979.