Universal Journal of Educational Research 8(3): 895-903, 2020 http://www.hrpub.org
DOI: 10.13189/ujer.2020.080321

Capability Estimation of Student's Higher Order Thinking in Mathematics by Using Polytomous

M. Zainudin[1,2,*], Bambang Subali[3], Jailani[3]

[1]Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia
[2]Department of Mathematics Education, IKIP PGRI Bojonegoro, Indonesia
[3]Universitas Negeri Yogyakarta, Indonesia

Received December 6, 2019; Revised January 14, 2020; Accepted February 7, 2020

Abstract One of the alternative approaches in the scoring of items that can be used is the Latent Trait Theory Approach for polytomous scoring with a Generalized Partial Credit Model (GPCM). This study aimed at estimating the test taker's capability to think critically (higher order thinking) in maths in the form of essay questions with items scoring in GPCM by considering the level of difficulty in each of the steps that must be taken to obtain the correct answer. The instrument used in this study was essay question developed. The data analysis used in this study was SPSS version number 22 and Parscale from Ssi. SPSS version number 22 was used to detect the sufficiency requirements of the number of samples used and prerequisite analysis of scoring using the GPCM method. Parscale was used to estimate the capability of respondents. The estimation of capability was presented in Phase 3 on a scale (-4, +4) in which, then, it can be transformed with a linear transformation to the equation $x = 12{,}5\theta + 50$, x is the estimation of capability in the range of [0,100]. The result of capability estimation was in the range of [0,100] so that it was obtained the minimum capability estimation 49.36 and maximum 100.

Keywords Capability Estimation, Higher Order Thinking, Latent Trait Theory, Polytomous

1. Introduction

The development of the world in the era of globalization in line with the era of knowledge (knowledge age) is very quickly implicated in various areas of life, including education. To deal with that situation, education should be able to prepare a generation that has higher order thinking skills, so they are expected to think critically, research, solve problems, make decisions, and have good character. [1] states that in order to deal with the rapidly changing world, education in the century of knowledge (the twenty-first century) must be able to develop the habit of critical thinking, researching, and solving problems. The same thing was also conveyed by [2] that in order to deal with the rapid changes, learners need to be skilled about how to learn and how to think.

The mathematics that is a part of the knowledge which is learned in all education levels should be taught deeply and appropriately so that the learners have the capability to develop their own potential in making decisions and solving problems. One of the alternative approaches to mathematics learning that can develop the potential in making decisions and solving problems is Higher Order Thinking Skills (HOTS) approach. According to [3], by having HOTS, then someone will be able to learn, to give good reasoning, to think creatively, to make a decision, and to solve the problem.

Some of the above-mentioned capabilities can be achieved if one is able to apply science, analyze problems, evaluate problems, and develop an alternative solution based on their existing knowledge and understanding. Some of those capability indicators are summarized in HOTS, so HOTS must be owned by all students including students at Junior High School level.

[4] states that HOT requires students to manipulate information and ideas in ways that transform Reviews their meaning and implications, such as when students combine facts and ideas in order to synthesize, generalize, explain, hypothesize, or arrive at some conclusion Or interpretation. By applying HOT, students will learn more deeply, knowledge is thick. Students will understand the concept better. It deals with the substantive character for a lesson when learners are able to demonstrate their understanding well and deeply. By having HOT, learners can clearly distinguish ideas or concepts, argue well, able to solve problems, able to construct explanations, able to

formulate a hypothesis and understand complex things become clearer. [5] suggest that HOT can be learned, HOT can be taught to students, with HOT the skills and the character of learners can be improved.

Thinking means using such analytical, creative, practical, and intelligence skills in everyday life. Higher order thinking skills such as metacognitive are part of high-order thinking. [6] state that HOT means the capacity to go beyond the information given, to adapt metacognitive awareness and to solve problems.

One of the HOT-based instruments in the question is used in the Program for International Students Assessment (PISA). According to [7] who classifies the level of questions on the PISA with the level of thought according to Bloom, found that the level of 4 (four) to level 6 (six) questions on PISA are categorized into Higher Order Thinking, while level 1 (one) to level 3 (three) are categorized as Low Order Thinking. McMahon in [7] states that the Higher Order Thinking process is an integration of critical thinking processes and creative thinking processes. Based on [8], he argues that only 0.1 percent of students who are able to solve the problem until the fifth and the sixth levels. It means that the capability of Higher Order Thinking Skill of Indonesian students is still low, it also means that the critical thinking skills of Indonesian students still need to be optimized.

[8] argues that familiarizing HOT-based assignment to junior high school students to develop high-level thinking potential is needed. One of the HOT-based problems of mathematics can be arranged in the form of an essay.

Scoring on the item in the form of an essay can be done by looking at the stages of test takers in solving the problem. In the classical test theory approach, the calculation of the score is done by summing the overall score obtained by the students. This approach is not necessarily appropriate because the level of difficulty of each step is not taken into account. In addition, the probability of correct answer given by the students based on a particular response is unpredictable [9]. Therefore, alternative approaches are needed, including using Latent Trait Theory (LTT) for polytomous scoring, one of them is by using the generalized partial credit model (GPCM) which is an extension of the Partial Credit Model (PCM) [10].

In the early development of Latent Trait Theory (LTT) for polytomous scoring, the well-known model is the extension of *Rasch Model* which is called as the Partial Credit Model (PCM). PCM is a schematic model of polytomous which is an extension of *Rasch Model* on dichotomous data. The assumption on PCM that each item has the same discriminating power. The PCM has a similarity to the Graded Response Model (GRM) on the scored item in the gradual category, but the index of difficulty in each step does not need to be sorted, a step can be more difficult than the next step.

[11] state that the assumption must be met in the GPCM method is a single dimension (unidimensional) and local independence. The assumption of a single dimension can be checked through root characteristics in the factor analysis, calculating the ratio between the first and second roots. If the ratio is high, then the model is unidimensional. Local independence is the response of the test takers to an item not related to any other item in the test. The IRT frees the test takers and the items of interdependence so that the difficulty level of the item is no longer dependent on the capability of the test takers, the capability of the test takers no longer depends on the difficulty of the item.

GPCM is formulated on the assumption that every probability of choosing the k-category beyond the (k-1) category is built by the dichotomy model. P_{ik} is a special probability of choosing the k-th category of mj +1 category. The probability relationship is correctly answered for each capability θ presented in the Categorical Response Function (CRF) graph [14].

This study was conducted to estimate the capability of high-order thinking of the test takers in the form of an essay problem with the scoring of items on GPCM-based polytomous Latent Trait Theory.

2. Methods

This study uses data obtained from the results of answers to the problem of higher order thinking in the form of an essay of the problem adapted from the PISA problem on the connection and reflection process competencies developed and has proven valid and reliable with the reliability index of 0.743. The number of items in the question used in this study is 5 items. Implementation of the test was held on June 3, 2017, at all Junior High Schools in Balen Subdistrict, Bojonegoro, East Java, Indonesia. The number of test takers is 315 respondents.

The data analysis to prove the assumption of Item Response Theory (IRT) or Latent Trait Theory (LTT) was done by using the SPSS version number 22 program on the unidimensional and local independence tests, as well as the Bilog-MG program for parameter invariance tests. The program used to estimate item parameters and capabilities in this study was Parscale of Ssi [12]. Two things that need to be concerned with the use of Parscale program from Ssi are data input and syntax analysis. Data Input can use * .txt text format, similarly, the syntax analysis. The software used in this study was Parscale software from Ssi.

3. Results and Discussion

[11] state that the underlying assumptions of GPCM are unidimensional and local independence. Unidimensional

means that each of test item measures only one capability. Unidimensional assumptions can only be shown only if the test contains a dominant component that measures the subject's capability. Unidimensional assumptions can be proven by using factor analysis, to look at the eigen values in the variance matrix of inter-item covariance.

The analysis of sample adequacy with SPSS version number 22 shows the value of chi-square on Bartlet test is 1360.219 with the degree of freedom is 28 and p-value are less than 0.01. The results show that the sample size of 315 used in this study meets the prevailing norms and is also corroborated with the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) of 0.947 which is greater than 0.5. The full results are presented in Table 1.

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.947
Bartlett's Test of Sphericity	Approx. Chi-Square	1360.219
	Df	28
	Sig.	.000

Unidimensional assumptions can be determined based on the scree plot or can be determined by looking at a 20% variance percentage or greater by 5 or 4 [20]. Based on data analysis with SPSS version number 22 program obtained the following results.

Table 2. Total Variance Explained

factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.979	62.234	62.234	4.555	56.942	56.942
2	.547	6.843	69.077			
3	.527	6.586	75.663			
4	.440	5.503	81.166			
5	.419	5.244	86.410			
6	.394	4.922	91.332			
7	.366	4.574	95.906			
8	.328	4.094	100.000			

Extraction Method: Principal Axis Factoring.

While the scree plot results of data analysis of student responses to mathematics higher order thinking with generalized partial credit model using SPSS version number 22 is as follows.

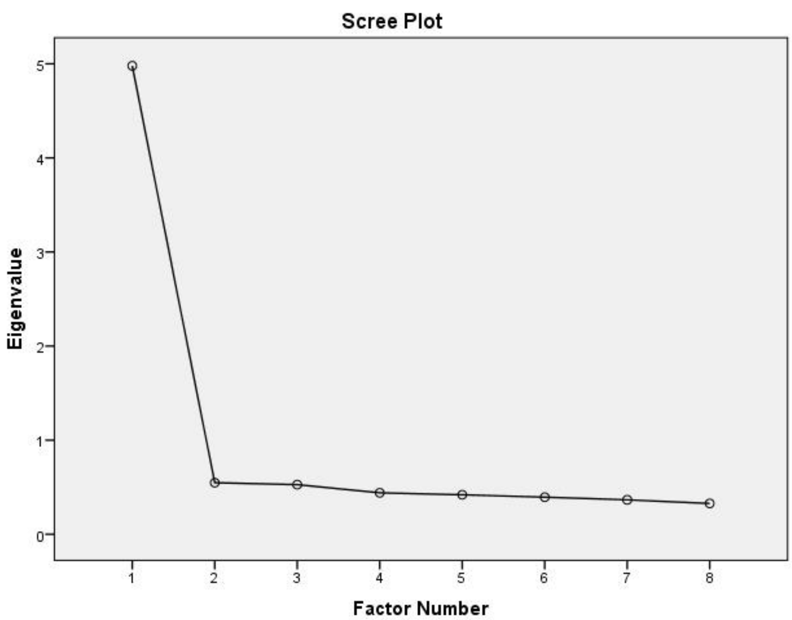

Figure 1. Plot of Exploratory Factor Analysis Result

Based on Table 2 and the scree plot shown in Figure 1 it can be concluded that the Eigen values and variance components of factor analysis by using SPSS version number 22 from the student response data to the mathematics higher order thinking with the generalized partial credit model contain 1 dominant factor , So the unidimensional assumption is met.

The assumption of local independence is divided into two: local independence of the test takers' responses and local independence on the test items (James J. Allen & Yen, 1989: 241). The local independence of the test takers' responses means that the test takers' correctness or wrong answer to an item is not affected by the correctness or wrong of the other test takers in answering the item. While local independence to the items means that correctness or wrong of the test takers answered an item is not affected by the correctness or wrong of the test takers in answering another item. Local independence is a condition where if the factors affecting achievement are constant, then the response of the subject to any pair of items will be statistically independent of each other. This local independence assumption will be fulfilled if the participant's response to an item does not affect the participant's response to another item. This assumption can be proven because the unidimensionality of the participant's response data to a test is met. The test to fulfill the assumption of local independence can be done by proving that the chances of the answer pattern of each test participant are equal to the result of the chance of the test takers' answer on each item.

The two underlying assumptions of GPCM, i.e. unidimensional and local independence are met, so it can be analyzed by Parscale program to estimate the capability of test takers. The results of data analysis of higher order thinking of mathematics by using generalized partial credit model based polytomous latent trait theory with the help of parscale program from Ssi are as follows.

```
>COMMENTS

>FILES    DFNAME='zaynn.TXT', SAVE;
>SAVE     PARM='zaynn.PAR',SCORE='zaynn.SCO';
>INPUT    NIDW=3, NTOTAL=8, NTEST=1, LENGTH=8;
(3A1,2X,8A1)
>TEST     TNAME='SCALE1', ITEM=(1(1)8), NBLOCK=2;
>BLOCK1   BNAME='SBLOCK1', NITEMS=4, NCAT=4, SCORING=(1,2,3,4);
>BLOCK2   BNAME='SBLOCK2', NITEMS=4, NCAT=4, MODIFIED=(1,1,2,2), SCOR=(1,2);
>CALIB    PARTIAL, LOGISTIC, NQPTS=15, CYCLE=(100,1,1,1,1), NEWTON=2,
             CRIT=0.01;
>SCORE    MLE, SMEAN=0.0, SSD=1.0, NAME='PCR_MLE', PFQ=5;
```

Figure 2. Syntax Analysis for GPCM with Parscale of Ssi

Once done running from Parscale, it was obtained the output in the form of *.Ph1 (phase 1), * .Ph2 (phase 2), and *.Ph3 (phase 3). Phase 1 is the result of parameter estimation based on the classical test theory, which includes the proportion of true answer step and the polyserial correlation which is the item of discriminating power. Phase 2 is the estimation phase of the item parameters, and phase 3 is the estimation of test takers' capability. The results of each phase are shown in Figures 3, 4, and 5.

```
BLOCK NO.:   1      NAME: SBLOCK1
--------------------------------------------------------------
ITEM    | TOTAL    NOT      OMIT |          CATEGORIES
        |        PRESENT         |
        |                        |  1       2       3       4
--------------------------------------------------------------
0001    |                        |
  FREQ.|   315       0         0|   67      98      92      58
  PERC.|           0.0       0.0|  21.3    31.1    29.2    18.4
        |                        |
0002    |                        |
  FREQ.|   315       0         0|   65      83     111      56
  PERC.|           0.0       0.0|  20.6    26.3    35.2    17.8
        |                        |
0003    |                        |
  FREQ.|   315       0         0|   80      85      93      57
  PERC.|           0.0       0.0|  25.4    27.0    29.5    18.1
        |                        |
0004    |                        |
  FREQ.|   315       0         0|   57     102     101      55
  PERC.|           0.0       0.0|  18.1    32.4    32.1    17.5
        |                        |
--------------------------------------------------------------
CUMMUL.|                        |
  FREQ.|                        |  269     368     397     226
  PERC.|                        |  21.3    29.2    31.5    17.9
--------------------------------------------------------------
```

```
BLOCK NO.:   2     NAME: SBLOCK2
-----------------------------------------------
ITEM      | TOTAL     NOT      OMIT |  CATEGORIES
          |         PRESENT         |
          |                         |  1       2
-----------------------------------------------
0005      |                         |
  FREQ.|     315        0         0|   163     152
  PERC.|              0.0       0.0|  51.7    48.3
          |                         |
0006      |                         |
  FREQ.|     315        0         0|   207     108
  PERC.|              0.0       0.0|  65.7    34.3
          |                         |
0007      |                         |
  FREQ.|     315        0         0|   220      95
  PERC.|              0.0       0.0|  69.8    30.2
          |                         |
0008      |                         |
  FREQ.|     315        0         0|   207     108
  PERC.|              0.0       0.0|  65.7    34.3
          |                         |
-----------------------------------------------
CUMMUL.|                            |
  FREQ.|                            |   797     463
  PERC.|                            |  63.3    36.7
-----------------------------------------------
```

```
-------------------------------------------------------------------------------
BLOCK      |  RESPONSE    TOTAL SCORE | PEARSON  &  |  INITIAL      INITIAL
  ITEM     |   MEAN          MEAN     | POLYSERIAL  |   SLOPE       LOCATION
           |   S.D.*         S.D.*    | CORRELATION |
-------------------------------------------------------------------------------
 SBLOCK1   |                          |             |
  1 0001   |    2.448       15.311    |    0.805    |     1.684      -0.005
           |    1.020*       4.591*   |    0.860    |
  2 0002   |    2.502       15.311    |    0.809    |     1.728      -0.006
           |    1.009*       4.591*   |    0.866    |
  3 0003   |    2.403       15.311    |    0.837    |     2.043       0.045
           |    1.054*       4.591*   |    0.898    |
  4 0004   |    2.489       15.311    |    0.832    |     1.925      -0.016
           |    0.980*       4.591*   |    0.887    |
-------------------------------------------------------------------------------
 CATEGORY  |  SCORING  |     MEAN     |     S.D.    | PARAMETER
    1      |    1.000  |    10.089    |    2.140    |
    2      |    2.000  |    13.454    |    2.791    |     0.852
    3      |    3.000  |    17.219    |    2.907    |     0.037
    4      |    4.000  |    21.199    |    2.323    |    -0.889
-------------------------------------------------------------------------------
 SBLOCK2   |                          |             |
  5 0005   |    1.483       15.311    |    0.747    |     2.667      -0.188
           |    0.500*       4.591*   |    0.936    |
  6 0006   |    1.343       15.311    |    0.639    |     1.458       0.073
           |    0.475*       4.591*   |    0.825    |
  7 0007   |    1.302       15.311    |    0.585    |     1.209       0.225
           |    0.459*       4.591*   |    0.771    |
  8 0008   |    1.343       15.311    |    0.655    |     1.583       0.051
           |    0.475*       4.591*   |    0.845    |
-------------------------------------------------------------------------------
 CATEGORY  |  SCORING  |     MEAN     |     S.D.    | PARAMETER
    1      |    1.000  |    13.033    |    3.601    |
    2      |    2.000  |    19.233    |    3.275    |     0.000
-------------------------------------------------------------------------------
```

Figure 3. Output Phase 1

Based on Figure 1 it can be concluded that each item has a category percentage of more than 5%, so there is no need to merge categories that have less than 5% presentation with another category which has presentation category more than 5%. Based on Figure 3, it is revealed the result of parameter estimation based on classical test theory and polyserial correlation which is the item of discriminating power.

```
ITEM BLOCK   1  SBLOCK1

SCORING FUNCTION   :    1.000     2.000     3.000     4.000
STEP PARAMTER      :    0.000     0.929     0.069    -0.998
S.E.               :    0.000     0.049     0.043     0.051
ITEM BLOCK   2  SBLOCK2

SCORING FUNCTION   :    1.000     2.000
STEP PARAMTER      :    0.000     0.000
S.E.               :    0.000     0.000
+------+-----+---------+---------+---------+---------+---------+---------+
| ITEM |BLOCK|  SLOPE  |  S.E.   |LOCATION |  S.E.   |GUESSING |  S.E.   |
+======+=====+=========+=========+=========+=========+=========+=========+
| 0001 |   1 |   0.975 |   0.118 |   0.073 |   0.077 |   0.000 |   0.000 |
| 0002 |   1 |   1.073 |   0.122 |   0.001 |   0.076 |   0.000 |   0.000 |
| 0003 |   1 |   1.091 |   0.139 |   0.129 |   0.072 |   0.000 |   0.000 |
| 0004 |   1 |   1.324 |   0.191 |   0.017 |   0.073 |   0.000 |   0.000 |
+------+-----+---------+---------+---------+---------+---------+---------+
| 0005 |   2 |   1.963 |   0.292 |   0.044 |   0.081 |   0.000 |   0.000 |
| 0006 |   2 |   1.220 |   0.168 |   0.519 |   0.099 |   0.000 |   0.000 |
| 0007 |   2 |   0.988 |   0.151 |   0.733 |   0.118 |   0.000 |   0.000 |
| 0008 |   2 |   1.290 |   0.193 |   0.507 |   0.095 |   0.000 |   0.000 |
+------+-----+---------+---------+---------+---------+---------+---------+

 SUMMARY STATISTICS OF PARAMETER ESTIMATES

   +----------+---------+---------+----+
   |PARAMETER |   MEAN  | STN DEV | N  |
   +==========+=========+=========+====+
   |SLOPE     |    1.241|    0.320|   8|
   |LOG(SLOPE)|    0.191|    0.226|   8|
   |THRESHOLD |    0.253|    0.287|   8|
   |GUESSING  |    0.000|    0.000|   0|
   +----------+---------+---------+----+
```

Figure 4. Output Phase 2

```
SUBJECT IDENTIFICATION                 WEIGHT/FREQUENCY
SCORE NAME    GROUP    WEIGHT MEAN CATEGORY ATTEMPTS    ABILITY       S.E.

--------------------------------------------------------------------------------
  1                |         1  GROUP 01      1.00
  1  PCR_MLE     1 |      1.00        2.50    1.00        0.9839      0.2897
--------------------------------------------------------------------------------
  2                |         2  GROUP 01      1.00
  1  PCR_MLE     1 |      1.00        1.38    1.00       -0.9266      0.3180
--------------------------------------------------------------------------------
  3                |         3  GROUP 01      1.00
  1  PCR_MLE     1 |      1.00        1.50    1.00       -0.7762      0.3009
--------------------------------------------------------------------------------
  4                |         4  GROUP 01      1.00
  1  PCR_MLE     1 |      1.00        1.50    1.00       -0.6784      0.2909
--------------------------------------------------------------------------------
```

Figure 5. Parameter Estimation Output (* .PAR)

Capability estimation results are presented at the interval scale (-4, +4). In order to be used for better interpretation, it is necessary to do with linear transformation with the equation x = 12,5θ + 50, for x is a capability estimation in the range [0,100]. The result of capability estimation in the range [0,100], so that it was obtained the minimum capability estimation 49,36 and maximum 100.

```
File  Edit  Format  View  Help
EXAMPL03.PSL - ARTIFICIAL EXAMPLE (MONTE CARLO DATA)
              GENERALIZED PARTIAL CREDIT MODEL - EAP SCALE SCORES
TEST  01    2    8    7    0    1
    4    4
GROUP 01
SBLOCK1     40001   0.97546   0.11775   0.07279   0.07713   0.00000   0.00000
SBLOCK1     40002   1.07323   0.12238   0.00136   0.07598   0.00000   0.00000
SBLOCK1     40003   1.09093   0.13875   0.12855   0.07228   0.00000   0.00000
SBLOCK1     40004   1.32354   0.19062   0.01734   0.07348   0.00000   0.00000
   0.00000   0.92908   0.06919  -0.99827
   0.00000   0.04891   0.04305   0.05071
SBLOCK2     20005   1.96309   0.29229   0.04407   0.08072   0.00000   0.00000
SBLOCK2     20006   1.22017   0.16794   0.51852   0.09862   0.00000   0.00000
SBLOCK2     20007   0.98814   0.15087   0.73282   0.11798   0.00000   0.00000
SBLOCK2     20008   1.29040   0.19251   0.50726   0.09482   0.00000   0.00000
   0.00000   0.00000
   0.00000   0.00000
```

Figure 6. Output Phase 3

The result of the average score with the capability can be shown in the following plot.

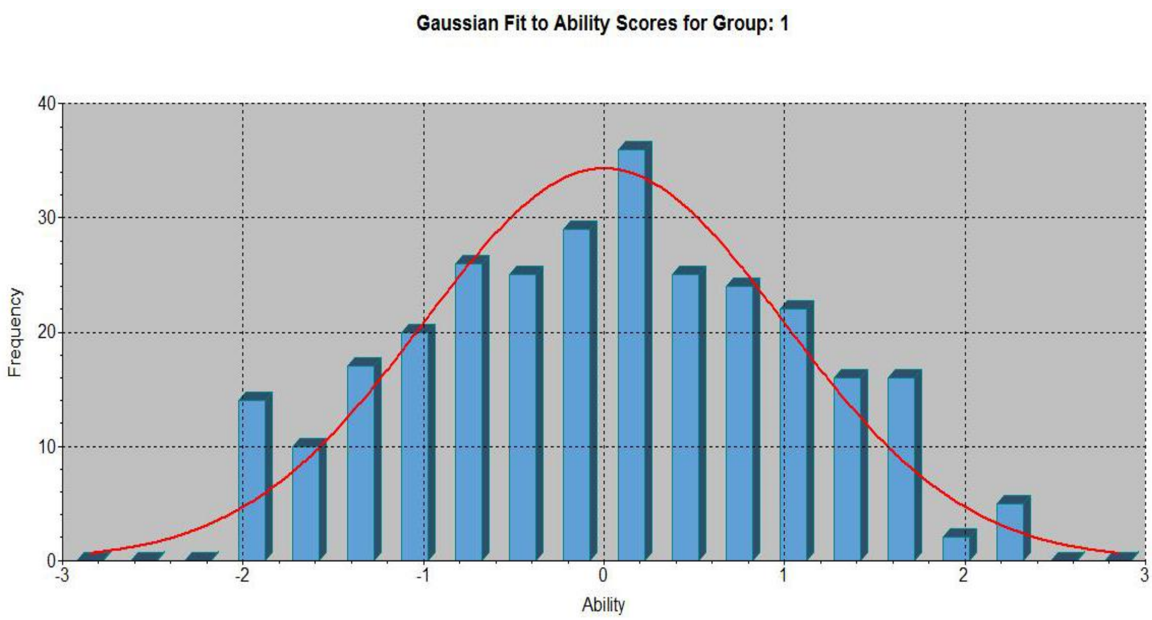

Figure 7. The plot of mean score with capability

Based on Figure 7 it can be concluded that the average score with the capability to solve the problem of high-order thinking of mathematics forms a normal curve. Respondents who have average capabilities are more than those with low and high capability.

The results of the study show the tendency of students' high order thinking skills. High order thinking skills spread to form normal curves. Analytical skills are related to the ability to identify the main ideas of problems, analyze arguments, and compare and contrast the previous knowledge.

Students who have an average score can identify the main idea by stating what is known and asked on the question, concisely, and precisely for all questions. Next, the student gives a theoretical reason in each step of the work up to the final answer correctly for several questions.

The student is also able to provide similarities, differences, and uses of things that are known to answer the questions correctly for some questions.

The ability of evaluation is related to the ability to provide an assessment of the solutions and methods used in answering questions and criticize arguments. Students who have an average score can provide an assessment of the solutions and methods used in answering questions correctly for several questions. The student is confident in the answers and the methods used in answering some questions. It is based on logical and theoretical answers and the methods used. Next, the student can criticize the argument appropriately for several questions. Students do a double check starting from what is known to the conclusion of the answer by paying attention to the theoretical aspects of the work steps correctly for several questions.

Creative ability relates to the ability to design ways of solving problems and making new steps. Students who have an average score can design ways of solving problems to answer several questions correctly. Students have to design a method by considering preliminary analysis on things that are known and asked on the questions so that the correct way to answer the question is obtained. Furthermore, the student can make new steps by combining the steps of the previous work logically and theoretically for several questions.

Logical and reasoning abilities include the content of answers, reasoning, and evidence, as well as clarity of language style. Students who have an average score have complete, systematic, and theoretical steps for answers to several questions. Next, the student gives the reason for the work logically and writes the answers clearly and effectively.

In this study, it can be seen the relationship between the ability of creation with the ability of analysis and evaluation. Creative ability will not be able to stand alone, in the sense of creative ability is influenced by the ability of analysis and evaluation. Students who have an average score on the aspects of analysis and evaluation, resulting in a moderate level of creativity as well. This also applies to students whose analysis and evaluation skills are below average; they tend to have below average creative abilities as well. The results of the above research are relevant to the result of [21], which states that the students can create something if they have been able to do an analysis and evaluation first.

4. Conclusions

The scoring on a mathematics higher order thinking skills test is done with polytomous with the Generalized Partial Credit Model (GPCM). Capability estimation can be done with the help of Parscale of SSi. Capability estimation was presented in Phase 3 on a scale (-4, +4) which then can be transformed further with a linear transformation to the equation $x = 12{,}5\theta + 50$, x is the estimation of the capability in the range [0,100]. The result of capability estimation was in the range [0,100] so that it was obtained the minimum capability estimation 49.36 and maximum 100.

Based on the results of this study, deepening of the capability estimation can be done by using simulation data between using the correct answer score and using GPCM.

The suggestions for further researchers is that this research can be used as input for conducting similar research both in developing instruments to improve students' high order thinking skills, as well as analyzing the factors that affect these abilities. Besides, to be more valid in measuring students' high order thinking skills, high order thinking skills validity and reliability tests should be carried out

REFERENCES

[1] D. R, How to use problem-based learning in the classroom. Alexandria: ASCD, 1997.

[2] C. Rose and M. J. Nicholl, Accelerated learning for the 21st century. Bandung: Nuansa, 2000.

[3] J. P. Robinson, “What are employability skills?,” Work., vol. 1, no. 3, pp. 27–38, 2000.

[4] F. M. Newman and G. G. Wehlage, “Five standards of authentic instruction,” J. Dep. Superv. Curric. Dev., vol. 50, no. 7, 2011.

[5] A. Thomas and G. Thorne, “Higher level thinking-it’s HOT!,” the center for development and learning, 2009. .

[6] C. Mcloughlin, “An e-learning solution to creating work-related skills and competencies for the knowledge-based economy and competencies for the knowledge-based economy,” in Proceedings of the Apple University Consortium Conference, 2011, vol. 20, no. 2, pp. 14–19.

[7] H. Setiawan, “Math problems in PISA,” in National Conference of Mathematics, 2014.

[8] Stacey, The PISA of Mathematical Literacy in Indonesia. Australia: University of Melbourne, 2011.

[9] G. N. Masters and J. P. Keeves, Advances in measurement in educational research and assesment. Amsterdam: Pergamon, 1999.

[10] E. Muraki, “New approaches to measurement,” in Advances in measurement in educational research and assessment, J. P. (Eds. Masters, G.N. dan Keeves, Ed. Amsterdam: Pergamon, 1999.

[11] H. Hambleton, R.K. Swaminathan, Item response theory. Boston, MA: Kluwer Inc, 1985.

[12] E. Muraki and R. D. Bock, Parscale 3: IRT based test scoring and item analysis for graded items and rating scales. Chicago: Scintific Software Inc, 1997.

[13] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamental of item response theory. Newbury Park, CA: Sage Publication Inc, 1991.

[14] M. Du Toit, IRT from SSi: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood: SSi, 2003.

[15] M. S. Johnson, “Journal of Statistical Software,” J. Stat. Softw., vol. 20, no. 10, 2007.

[16] R. V Hogg and A. T. C. McKean, Introduction to mathematical statistics. Boston, MA: Pearson Education, Inc, 2013.

[17] T. Wang, “Relative precision of ability estimation in polytomous CAT: A comparisonunder the Generalized Partial Credit Model and Graded Response Model,” in the annual meeting of AERA Associations, 2002.

[18] F. B. Baker, The Basics of Item Response Theory. Washington, DC: ERIC Publications, 2001.

[19] W. W., Polythomous model in Item Respons Theory. Yogyakarta: UGM, 2010.

[20] C. S. Wells, R. K. Hambleton, and U. Purwono, “Item response theory: polytomous respons IRT models and application,” in Workshop of educational assessment and psicometry, 2008.

[21] Kurniati, D., Harimukti, R. Jamil, N. A. The Junior High School Students’ Ability to Use High Order Thinking Skills in Solving Pisa-Based Problems. Jurnal Penelitian dan Evaluasi Pendidikan [Journal of Educational Research and Evaluation], 20 (2).142-155, 2016.