The New Educational Review

DOI: 10.15804/tner.2023.72.2.16

Indah Dwi Harvani
University of Jember, Indonesia

Dian Kurniati
University of Jember, Indonesia

Dong-Joong Kim
Korea University, South Korea

Sharifah Osman
Universiti Teknologi Malaysia, Malaysia

Quality of Algebraic Numeration Problems to Measure Higher Order Thinking Skills Using Partial Credit Model

Abstract

Higher-order thinking skills are abilities that all students must possess. Students are given algebraic numeracy questions to measure this ability. This research aims to determine the quality of algebraic numeracy questions based on their validity and reliability. This type of research is quantitative, with analysis using a partial credit model with 28 students as research subjects. The results in this study were that the 10 numeration questions developed were valid because they met at least one of the validity and reliability criteria with an item reliability value of 0.82, indicating that the instrument could measure high-level thinking skills well.

Keywords: *algebraic numeration problems, higher-order thinking skills, partial credit model*

Introduction

The main objective of implementing the independent learning program by the Minister of Education in Indonesia is to improve the quality of human resources through students who have noble character and can perform high-level reasoning,

especially in the areas of literacy and numeracy. To achieve this main goal, the Indonesian government has implemented a national assessment which has been implemented since 2021. This assessment evaluates the education system at the primary and secondary education levels. Furthermore, the Indonesian government responded to the results of the PISA analysis from 2003 to 2018, the average score in mathematics was 377. This score fluctuated in the early years of PISA but has been relatively stable since 2009 (OECD, 2019). It shows that Indonesian students' literacy and numeracy abilities must be improved. Efforts to improve core skills for measuring literacy and numeracy abilities are by establishing a minimum competency assessment.

The minimum competency assessment questions require students to use various cognitive skills. The minimum competencies that should be mastered are mathematical reasoning abilities and critical reading skills. In addition, other abilities, such as interpreting, integrating, and evaluating information in various forms, also need to be possessed by students. To arrive at the evaluating and reflective level, students must first have the ability to think creatively and critically. In line with this, numeration questions must be a matter of higher-order thinking skills (Sani, 2021).

According to Mullis & Martin (2017), the cognitive level of the minimum competency assessment numeration questions based on the International Association for the Evaluation of Educational Achievement is divided into knowing, applying, and reasoning levels (Assessment and Learning Center, 2020). The knowing level belongs to the High Order Thinking Skill (HOTS) level, namely analysing and evaluating. Meanwhile, the level of applying and reasoning belongs to the High Order Thinking Skill (HOTS) level, namely analysing, evaluating, and creating. HOTS-type questions have been introduced in the exam, and integration into them is expected to increase in 2020 (Zaharin et al., 2018).

Thinking skills are intellectual processes that involve forming concepts, applications, analysis, syntax, and evaluating information collected based on experience and observation (Yee et al., 2015). Through Bloom's taxonomy model, thinking skills form the cognitive domain of remembering, understanding, applying, analysing, evaluating, and creating. It is divided into two levels of thinking, namely, high-order thinking skills and low-order thinking skills. Each level is connected where students can get the highest level of creative skills if they can remember and understand the concept as a whole and apply their understanding by analysing, evaluating, and creating (Zaharin et al., 2018).

The evaluation process is carried out to determine the quality of the instrument. The algebraic numeracy questions in the minimum competency assessment are

tests designed to determine higher-order thinking skills. The results need to be analysed to determine deficiencies to correct them. The evaluation process is carried out using the Rasch model. The main objective is to produce measurement parameters at the same intervals to obtain accurate information regarding the research subject and the quality of algebraic questions in the minimum competency assessment being tested. The forms of questions in the numeration assessment test are quite diverse, namely multiple choice, short essays, and essays, so the maximum score for each item is also different. In Rasch modelling, to analyse data with different maximum scores, you can use the Partial Credit Model (Wahyuningsih, 2021).

Research Methodology

Participants and Procedure

Algebraic numeration questions were tested on 28 research subjects. The research subjects were chosen because they implemented an independent curriculum and completed a Minimum Competency Assessment. The selected research subjects have various higher-order thinking abilities.

The first step to using the Partial Credit Model (PCM) is a value summary. Prepare raw data from the results of testing algebraic numeracy questions with different maximum scores. Then the raw data from Excel will be formatted into .prn. Furthermore, the data will be analysed for the Ministep application. The final step is to interpret the quality of the algebraic numeration instrument.

Measure

Items that do not fit can be checked by looking at the following criteria: (1) *MNSQ*: 0,5<*MNSQ*<*1,5*; (2) *ZSTD*: -2<ZSTD<2; (3) *Pt Mean Corr*: 0,4<Pt Mean Corr<0,85 (Perez-Marmol & Brown, 2018; Wahyuningsih, 2020).

Statistical Analysis

Rasch modelling mainly focuses on item assessment and item checking efforts and helps identify measurement error sources (Abdellatif, 2023). The performance of the test items was evaluated based on reliability, validity, descriptive statistics, and separation index. The value of Cronbach's alpha, item reliability, and person

reliability indicates the quality of the instrument's reliability. Outfit ZSTD, Outfit MNSQ, and Pt Mean Coor scores will indicate the validity of the test instrument. To find out the results of measuring higher-order thinking skills can be seen through the value of separation.

Results and Discussion

Validity Analysis

To find out the validity of an item can be seen from three criteria, namely Pt Mean Corr, Outfit MNSQ, and Outfit ZSTD. The question can be declared valid if one of the criteria is met. Figure 1 will contain information regarding the fit of an item. Not all items meet the three criteria for an item to fit.

```
TABLE 10.1 E:\SKRIPSI\Uji Coba Rasch\PCM.prn     ZOU171WS.TXT  Mar 15 2023 23:52
INPUT: 28 Person  10 Item  REPORTED: 28 Person  10 Item  5 CATS MINISTEP 5.3.2.0
-------------------------------------------------------------------------------
Person: REAL SEP.: .86  REL.: .42 ... Item: REAL SEP.: 2.00  REL.: .80

         Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item  G |
|------------------------------------+----------+----------+-----------+-----------+--------|
|     2     27     28   -1.98    1.06|1.22   .53|1.88   .97|A .00   .23| 96.4  96.4| S2    B |
|     4     47     28    -.39     .42|1.47  1.48|1.34   .85|B .49   .48| 64.3  73.1| S4    A |
|     1     27     28   -1.98    1.06|1.14   .45| .66   .20|C .18   .23| 96.4  96.4| S1    B |
|     3     27     28   -1.98    1.06|1.14   .45| .66   .20|D .18   .23| 96.4  96.4| S3    B |
|     6     26     28    2.18     .33|1.09   .44|1.07   .36|E .42   .55| 57.1  57.3| S6    A |
|     7     23     28     .07     .54|1.01   .14|1.04   .28|e .36   .39| 85.7  84.0| S7    B |
|     5     25     28    2.28     .33| .76 -1.03| .98  -.01|d .54   .55| 60.7  58.7| S5    A |
|     8     23     28     .07     .54| .92  -.15| .77  -.21|c .46   .39| 85.7  84.0| S8    B |
|    10     24     28    2.39     .33| .82  -.75| .80  -.78|b .65   .55| 57.1  59.2| S10   A |
|     9     25     28    -.65     .66| .71  -.54| .37  -.65|a .57   .34| 92.9  90.1| S9    B |
|------------------------------------+----------+----------+-----------+-----------+--------|
| MEAN    27.4   28.0     .00     .63|1.03   .10| .96   .12|           | 79.3  79.6|        |
| P.SD     6.7     .0    1.67     .30| .22   .70| .40   .54|           | 16.4  15.4|        |
-------------------------------------------------------------------------------------------
```

Figure 1. Suitability of Question Items

Here is the description:

1. The S2 item only meets the ZSTD Outfit criteria.
2. Items S1, S3, and S7 only meet the Outfit MNSQ and Outfit ZSTD criteria.
3. In item S9, it only meets the Outfit ZSTD and Pt Mean Corr criteria.

4. The items S4, S5, S6, S8, and S10 meet the criteria for Outfit MNSQ, Outfit ZSTD, and Pt Mean Corr.

Based on the analysis above, five items meet the three validity criteria. While the other items only meet one or two criteria. However, this item is still considered fit or appropriate because it meets one criterion. So that it can be stated that all items are categorised as fit questions and no questions need to be changed or replaced.

Reliability Analysis

The results of testing the algebraic numeracy instrument from 28 research subjects who completed 10 questions are shown in the statistical summary in Figure 2. Cronbach's alpha value is 0.47 indicates that the reliability measure, namely the

```
TABLE 3.1 E:\SKRIPSI\Uji Coba Rasch\PCM.prn      ZOU171WS.TXT  Mar 15 2023 23:52
INPUT: 28 Person  10 Item  REPORTED: 28 Person  10 Item  5 CATS MINISTEP 5.3.2.0
-------------------------------------------------------------------------------

     SUMMARY OF 28 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                        MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       9.8      10.0        1.95     .80       .96   -.06    .96    .19 |
|  SEM        .4        .0         .22     .02       .11    .20    .20    .13 |
| P.SD       2.0        .0        1.16     .10       .55   1.06   1.02    .69 |
| S.SD       2.0        .0        1.18     .10       .56   1.08   1.04    .70 |
| MAX.      13.0      10.0        4.35    1.13      2.54   1.88   5.23   2.11 |
| MIN.       4.0      10.0       -1.11     .70       .05  -2.62    .05   -.86 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .88 TRUE SD     .75  SEPARATION   .86  Person RELIABILITY  .42 |
|MODEL RMSE    .80 TRUE SD     .84  SEPARATION  1.04  Person RELIABILITY  .52 |
| S.E. OF Person MEAN = .22                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .47  SEM = 1.43
STANDARDIZED (50 ITEM) RELIABILITY = .84

     SUMMARY OF 10 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                        MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      27.4      28.0         .00     .63      1.03    .10    .96    .12 |
|  SEM       2.2        .0         .56     .10       .07    .23    .13    .18 |
| P.SD       6.7        .0        1.67     .30       .22    .70    .40    .54 |
| S.SD       7.1        .0        1.76     .32       .23    .74    .42    .57 |
| MAX.      47.0      28.0        2.39    1.06      1.47   1.48   1.88    .97 |
| MIN.      23.0      28.0       -1.98     .33       .71  -1.03    .37   -.78 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .75 TRUE SD    1.50  SEPARATION  2.00  Item   RELIABILITY  .80 |
|MODEL RMSE    .70 TRUE SD    1.52  SEPARATION  2.17  Item   RELIABILITY  .82 |
| S.E. OF Item MEAN = .56                                                     |
-------------------------------------------------------------------------------
```

Figure 2. Statistics Summary

interaction between the subject and the item, is of poor quality. The person reliability value of 0.52 means that the stability of the answers from the research subjects is still weak. Of the 28 subjects, the error in answering the questions lies in the relatively diverse items for the same number. While the stability of the numeration questions can be seen from the item reliability value, which equals 0.82. This value indicates good-quality questions. Therefore, the questions developed are suitable for measuring the subject's ability.

Difficulty Level Analysis

The logit value for each question has been sorted from the highest to the lowest, presented in Figure 3. Item S10 has the highest logit value of +2.39, indicating that question S10 is the most difficult problem for the research subject. While the S1, S2, and S3 have the same logit value, and the lowest is -1.98, it means that the subject answers the easiest questions.

Figure 4 shows the distribution of scores on item S10. Subjects who obtained a maximum score of 2 were only 5 subjects out of a total of 28 subjects, or around 18%. There were 14 subjects, or 50% of the subjects who got a score of 1. While the rest, or 32%, were not able to solve this problem. To determine the difficulty level of the questions, the person logit value of +1.95 can be used as a reference as the

TABLE 13.1 E:\SKRIPSI\Uji Coba Rasch\PCM.prn ZOU171WS.TXT Mar 15 2023 23:52
INPUT: 28 Person 10 Item REPORTED: 28 Person 10 Item 5 CATS MINISTEP 5.3.2.0

Person: REAL SEP.: .86 REL.: .42 ... Item: REAL SEP.: 2.00 REL.: .80

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
10	24	28	2.39	.33	.82	-.75	.80	-.78	.65	.55	57.1	59.2	S10	A
5	25	28	2.28	.33	.76	-1.03	.98	-.01	.54	.55	60.7	58.7	S5	A
6	26	28	2.18	.33	1.09	.44	1.07	.36	.42	.55	57.1	57.3	S6	A
7	23	28	.07	.54	1.01	.14	1.04	.28	.36	.39	85.7	84.0	S7	B
8	23	28	.07	.54	.92	-.15	.77	-.21	.46	.39	85.7	84.0	S8	B
4	47	28	-.39	.42	1.47	1.48	1.34	.85	.49	.48	64.3	73.1	S4	A
9	25	28	-.65	.66	.71	-.54	.37	-.65	.57	.34	92.9	90.1	S9	B
1	27	28	-1.98	1.06	1.14	.45	.66	.20	.18	.23	96.4	96.4	S1	B
2	27	28	-1.98	1.06	1.22	.53	1.88	.97	.00	.23	96.4	96.4	S2	B
3	27	28	-1.98	1.06	1.14	.45	.66	.20	.18	.23	96.4	96.4	S3	B
MEAN	27.4	28.0	.00	.63	1.03	.10	.96	.12			79.3	79.6		
P.SD	6.7	.0	1.67	.30	.22	.70	.40	.54			16.4	15.4		

Figure 3. Item Measure

TABLE 13.3 E:\SKRIPSI\Uji Coba Rasch\PCM.prn ZOU171WS.TXT Mar 15 2023 23:52
INPUT: 28 Person 10 Item REPORTED: 28 Person 10 Item 5 CATS MINISTEP 5.3.2.0

Item CATEGORY/OPTION/DISTRACTOR FREQUENCIES: MEASURE ORDER

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	ABILITY MEAN	P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.	Item
10	0	0	9	32	.95	.84	.30	.7	.7	-.59	S10
	1	1	14	50	2.18	.95	.26	1.0	.9	.20	
	2	2	5	18	3.10	.69	.35	.8	.8	.46	
5	0	0	7	25	1.36	.69	.28	1.0	1.0	-.30	S5
	1	1	17	61	1.81	1.08	.27	1.2	1.9	-.15	
	2	2	4	14	3.60	.44	.25	.5	.5	.58	
6	0	0	7	25	1.09	1.20	.49	1.1	1.1	-.43	S6
	1	1	16	57	2.15	1.04	.27	1.2	1.1	.19	
	2	2	5	18	2.53	.77	.38	1.2	1.3	.23	
7	0	0	5	18	1.05	1.30	.65	1.2	1.1	-.36	S7
	1	1	23	82	2.15	1.03	.22	.9	.9	.36	
8	0	0	5	18	.81	1.14	.57	.9	.8	-.46	S8
	1	1	23	82	2.20	1.01	.21	.9	.9	.46	
4	0	0	3	11	.05	.85	.60	1.0	.8	-.57	S4
	1	1	3	11	2.23	.82	.58	2.4	2.4	.08	
	2	2	22	79	2.17*	.99	.22	1.0	1.1	.36	
9	0	0	3	11	.05	.85	.60	.6	.3	-.57	S9
	1	1	25	89	2.18	.97	.20	.6	.8	.57	
1	0	0	1	4	.88	.00		2.1	.7	-.18	S1
	1	1	27	96	1.99	1.16	.23	1.2	1.1	.18	
2	0	0	1	4	1.96	.00		6.0	1.9	.00	S2
	1	1	27	96	1.95*	1.18	.23	1.2	1.1	.00	
3	0	0	1	4	.88	.00		2.1	.7	-.18	S3
	1	1	27	96	1.99	1.16	.23	1.2	1.1	.18	

Figure 4. Distribution Score

upper limit and the logit item of 0.00 as the lower limit. The results of the analysis of the difficulty level of the questions can be seen in Table 1.

Table 1. Level of Difficulty

Problem difficulty level	Item
Easy	S1, S2, S3, S4, and S9
Medium	S7 and S8
Difficult	S5, S6, and S10

Subject Ability Analysis

Figure 5 displays the measure column, which is the logit of each research subject. PJ05 has a logit value of +4.35. This value is the highest logit, meaning that the subject more often answers questions correctly compared to other subjects. Meanwhile, PJ13 has the smallest logit, -1.11, which indicates that the subject often answers the questions incorrectly. The average subject logit score is +1.95 (more than 0.00), meaning that subjects tend to answer the questions tested correctly.

TABLE 17.1 E:\SKRIPSI\Uji Coba Rasch\PCM.prn ZOU171WS.TXT Mar 15 2023 23:52
INPUT: 28 Person 10 Item REPORTED: 28 Person 10 Item 5 CATS MINISTEP 5.3.2.0

Person: REAL SEP.: .86 REL.: .42 ... Item: REAL SEP.: 2.00 REL.: .80

Person STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
5	13	10	4.35	1.13	.93	.17	.32	-.20	90.0	90.3	PJ05
4	12	10	3.35	.91	1.40	.77	1.48	.77	70.0	83.3	PJ04
12	12	10	3.35	.91	.63	-.44	.24	-.34	90.0	83.3	PJ12
16	12	10	3.35	.91	.57	-.57	.22	-.38	90.0	83.3	LJ16
18	12	10	3.35	.91	.57	-.57	.22	-.38	90.0	83.3	PJ18
27	12	10	3.35	.91	.51	-.71	.20	-.42	90.0	83.3	LJ27
7	11	10	2.60	.83	1.54	.92	.63	.17	80.0	84.7	PJ07
8	11	10	2.60	.83	.05	-2.62	.05	-.86	100.0	84.7	LJ08
10	11	10	2.60	.83	.05	-2.62	.05	-.86	100.0	84.7	LJ10
14	11	10	2.60	.83	2.54	1.88	2.04	1.04	60.0	84.7	LJ14
22	11	10	2.60	.83	1.54	.92	.63	.17	80.0	84.7	PJ22
1	10	10	1.96	.77	.53	-.76	.74	.23	90.0	82.0	LJ01
6	10	10	1.96	.77	.39	-1.17	.33	-.25	90.0	82.0	PJ06
15	10	10	1.96	.77	1.64	1.10	.78	.26	70.0	82.0	LJ15
17	10	10	1.96	.77	.53	-.76	.74	.23	90.0	82.0	LJ17
23	10	10	1.96	.77	.53	-.76	.74	.23	90.0	82.0	PJ23
24	10	10	1.96	.77	.67	-.44	5.23	2.11	90.0	82.0	PJ24
9	9	10	1.40	.73	1.61	1.19	.92	.24	60.0	76.1	LJ09
11	9	10	1.40	.73	.64	-.66	.40	-.48	80.0	76.1	PJ11
20	9	10	1.40	.73	.77	-.34	.65	-.07	80.0	76.1	LJ20
21	9	10	1.40	.73	.68	-.55	.42	-.44	80.0	76.1	LJ21
3	8	10	.88	.71	1.05	.26	2.18	1.44	80.0	70.6	PJ03
19	8	10	.88	.71	.96	.05	.89	.10	60.0	70.6	PJ19
25	8	10	.88	.71	1.56	1.24	1.02	.27	80.0	70.6	LJ25
28	8	10	.88	.71	1.31	.80	2.40	1.61	40.0	70.6	LJ28
2	7	10	.39	.70	1.34	.89	1.12	.40	40.0	71.7	LJ02
26	7	10	.39	.70	1.32	.85	.90	.02	80.0	71.7	PJ26
13	4	10	-1.11	.73	1.01	.16	1.28	.65	80.0	75.2	PJ13
MEAN	9.8	10.0	1.95	.80	.96	-.06	.96	.19	79.3	79.6	
P.SD	2.0	.0	1.16	.10	.55	1.06	1.02	.69	15.1	5.6	

Figure 5. Person Measure

Figure 5 obtained the value of separation = 1.04. Then the value of H = [(4 × 1.04) + 1]: 3 = 1.72, which, if rounded up, becomes 2. It means that the subjects are divided into two groups: the high-level group (subjects with many correct answers) and the low group (subjects with few correct answers). So based on the upper and lower limits of the logit value, the subjects can be divided into two groups with the same logit interval. Table 2 will display the frequency and percentage of each group.

Table 2. Frequency and Percentage of Subject Ability Groups

Ability Group (P)	Logit Intervals	Frequency	Percentage
Low Level Higher Order Thinking Skills	-1,11≤P<1,62	11	39%
High Level Higher Order Thinking Skills	1,62≤P≤4,35	17	61%

Figure 6 is a scalogram that will sort the subject's abilities from highest to lowest, while from left to right shows the subject's scores from the easiest to the most difficult questions.

```
TABLE 22.1 E:\SKRIPSI\Uji Coba Rasch\PCM.prn     ZOU171WS.TXT  Mar 15 2023 23:52
INPUT: 28 Person  10 Item  REPORTED: 28 Person  10 Item  5 CATS MINISTEP 5.3.2.0
-------------------------------------------------------------------------------

GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |         1
       |1239478650
       |----------
     5 +1111211122  PJ05
     4 +1111111221  PJ04
    12 +1111211112  PJ12
    16 +1111211121  LJ16
    18 +1111211121  PJ18
    27 +1111211211  LJ27
     7 +1111211012  PJ07
     8 +1111211111  LJ08
    10 +1111211111  LJ10
    14 +1111201202  LJ14
    22 +1111211012  PJ22
     1 +1111201111  LJ01
     6 +1111111111  PJ06
    15 +1111211200  LJ15
    17 +1111210111  LJ17
    23 +1111210111  PJ23
    24 +1011211111  PJ24
     9 +1111111200  LJ09
    11 +1111211100  PJ11
    20 +1111201110  LJ20
    21 +1111211010  LJ21
     3 +0111211100  PJ03
    19 +1110211010  PJ19
    25 +1111011101  LJ25
    28 +1101210011  LJ28
     2 +1110200011  LJ02
    26 +1111011100  PJ26
    13 +1110000010  PJ13
       |----------
       |         1
       |1239478650
```

Figure 6. Scalogram

A unique pattern of answers can be seen in LJ14 as a subject with high abilities. LJ14 was able to answer the most difficult questions with the maximum score but could not complete question number 7, which was in the medium category. PJ13 had low higher-order thinking ability, able to answer question number 5, which was in the difficult category but could not answer the other five questions with a lower level of difficulty. Seeing this kind of awkwardness, it is possible that the subject plagiarised his friend's answer or just guessed the answer, considering that question number 5 is a reasoning question that requires the subject's argument based on his calculations. According to (Nasrullah, 2022), students' ability to argue mathematically is supported by the creative motivation to logically and mathematically explain the solution to solving a given problem. The instruments tested are valid and reliable because they instrument tested are indicators of higher-order thinking skills (Septiani & Paidi, 2020).

Conclusions

The ten questions developed can be used as instruments to measure higher-order thinking skills because (1) they are valid, as seen from the MNSQ, ZSTD, and Pt. Mean Corr; (2) reliability is indicated by item reliability of 0.82 which indicates the numeracy questions are in a good category and can measure high-level thinking skills well; (3) effective because based on the analysis of student's abilities, as many as 40% of students are included in the low-level high-order thinking ability category, and the other 60% are included in the high-level high-order thinking ability category.

References

Abdellatif, H. (2023). Test Results with and without Blueprinting: Psychometric Analysis Using the Rasch Model. *Educación Médica, 24*, 1–14. https://doi.org/10.1016/j.edumed.2023.100802

Heong, Y. M., Yunos, J. M., Othoman, W., Hassan, R., Kiong, T. T., & Mohamad, M. M. (2012). The Needs Analysis of Learning Higher Order Thinking Skills for Generating Ideas. *Procedia-Social and Behavioral Sciences, 59*, 197–203. https://doi.org/10.1016/j.sbspro.2012.09.265

Hidayah, I. R., Kusmayadi, T. A., & Fitriana, L. (2021). Minimum Competency Assessment (AKM): An Effort to Photograph Numeracy. *Journal of Mathematics and Mathematics Education*, *11*(1), 14–20. https://doi.org/10.20961/jmme.v11i1.52742

Pérez-Mármol, J. M., & Brown, T. (2019). An Examination of the Structural Validity of the

Maslach Burnout Inventory-Student Survey (MBI-SS) Using the Rasch Measurement Model. *Health Professions Education, 5*(3), 259–274. DOI: 10.1016/j.hpe.2018.05.004

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218.

Kurniati, D., & Zayyadi, M. (2018). The Critical Thinking Dispositions of Students Around Coffee Plantation Area in Solving Algebraic Problems. *International Journal of Engineering and Technology (UAE), 7*(2), 18–20. https://doi.org/10.24127/ajpm.v11

Kwangmuang, P., Jarutkamolpong, S., Sangboonraung, W., & Daungtod, S. (2021). The Development of Learning Innovation to Enhance Higher Order Thinking Skills for Students in Thailand Junior High Schools. *Heliyon*, 7. https://doi.org/10.1016/j.heliyon.2021.e07309

OECD. (2019). Country Note: Programme for International Student Assessment (PISA) Results from PISA 2018. I-III: 1–10. https://www.oecd.org/pisa/publications/PISA2018_CN_IDN.pdf

OECD. (2019). PISA 2018 Results Combined Executive Summaries. I-III.

Perez-Marmol, J. M., & Brown, T. (2018). An Examination of the Structural Validity of the Maslach Burnout Inventory-Student (MBI-SS) Using the Rasch Measurement Model. *Health Professions Education, 5*(2), 259–274. https://doi.org/10.1016/j.hpe.2018.05.004

Sani, R. A. (2021). *Pembelajaran Berorientasi AKM*. Bumi Aksara.

Septiani, V., & Paidi. (2021). Development of A Test Instrument to Measure HOTS and Cognitive Knowledge in Biology. *Proceedings of the International Conference on Chemistry and Science Education (1-6)*. http://dx.doi.org/10.1088/1742-6596/1788/1/012028

Wahyuningsih, S., 2021. Using the Rasch's Partial Credit Model to Analyze the Quality of an Essay Math Test. *Proceedings of the 1st International Conference on Mathematics and Mathematics Education*. https://doi.org/10.2991/assehr.k.210508.073

Yee, M. H., Yunos, J. M., Othman, W., Hassan, R., Kiong, T. T., & Mohamad, M. M. (2015). The disparity of Learning Styles and Higher Order Thinking Skills Among Technical Students. *Procedia-Social and Behavioral Sciences,* 204, 143–152. http://dx.doi.org/10.1016/j.sbspro.2015.08.127

Zaharin, N. L., Sharif, S., & Mariappan, M. (2018). Computational Thinking: A Strategy for Developing Problem-Solving Skills and Higher Order Thinking Skills (HOTS). *International Journal of Academic Research in Business and Social Sciences, 8*(10), 1265–1278. http://dx.doi.org/10.6007/IJARBSS/v8-i10/5297

AUTHORS

INDAH DWI HARVANI

Undergraduate Degree (S.Pd), Student, University of Jember, Indonesia
E-mail: indahdwiharvani@gmail.com
ORCID: https://orcid.org/0009-0006-3164-5268

DIAN KURNIATI

Dr, Lecturer, University of Jember, Indonesia
E-mail: dian.kurniati@unej.ac.id
ORCID: https://orcid.org/0000-0002-9108-0434

DONG-JOONG KIM

Professor, Lecturer, Korea University, South Korea
E-mail: dongjoongkim@korea.ac.kr
ORCID: https://orcid.org/0000-0001-7607-5664

SHARIFAH OSMAN

Dr, Lecturer, Universiti Teknologi malaysia, Malaysia
E-mail: sharifah.o@utm.my
ORCID: https://orcid.org/0000-0003-2896-9377