Beta: Jurnal Tadris Matematika, 14(2) 2021: 126-141
DOI 10.20414/betajtm.v14i2.383

Research articles

Development and evaluation of a HOTS-based test for matrix topic: A classical test and item response theory

Muhamad Ali Misri[1], **Saifuddin**[1], **Reza Oktiana Akbar**[1], **Nok Rini Kamelia**[1]

Abstrak Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi soal tes berbasis keterampilan berpikir tingkat tinggi (HOTS) pada materi matriks. Pengembangan instrumen tes melalui dua tahap, yaitu pembentukan draf dan validasi. Pada tahap pertama, dilakukan kajian literatur yang relevan, penyusunan rencana butir soal, evaluasi butir soal yang diusulkan, dan uji coba draf butir soal. Sebanyak 51 siswa SMA dilibatkan pada tahapan uji coba. Pada tahap validasi, dilakukan analisis menggunakan teori tes klasik dan teori respon butir mencakup: karakterisasi, validitas dan reliabilitas, uji daya beda, dan tingkat kesulitan soal. Penelitian ini menghasilkan 5 butir soal yang valid (r_1=0,54; r_2=0,88; r_3=0,72; r_4=0,78; r_5=0,82). Tes yang dikembangkan mewakili materi matriks, memenuhi kriteria HOTS, dapat diandalkan dengan nilai reliabilitas tes sebesar r_α=0,85, dapat membedakan siswa yang memiliki kemampuan berpikir tingkat tinggi, dan memiliki keragaman tingkat kesulitan.

Kata kunci *Soal HOTS, Matriks, Teori klasik, Teori respon butir*

Abstract This research aims to develop and evaluate a higher-order thinking skills (HOTS)-based test for matrix topic. The development was carried out in two stages; items development and validation. The first stage was to review relevant literature about HOTS, design the test items, have experts' review, and tryout the items. Fifty-one upper secondary school students were involved in the tryout. In the second stage, results of the tryout were validated referring to the classical test and item response theory, including items characteristics, validity and reliability, items discrimination, and difficulty levels. The validation resulted in five valid test items (r_1=0,54; r_2=0,88; r_3=0,72; r_4=0,78; r_5=0,82). The developed test represents the topic, fulfils HOTS criteria, is reliable r_α=0,85, can differentiate students with higher-order thinking, and has varied difficulty levels.

Keywords *HOTS test, Matrix, Classic test theory, Item response theory*

Introduction

This study aims at developing a test instrument based on Higher Order Thinking Skills (henceforth: HOTS) on a matrix topic for secondary schools. This test is used to measure students' HOTS. Thus far, a test instrument is only used to confirm the teacher's explanation focusing on measuring students' knowledge (Zainudin, Subali & Jailani, 2019), nor on the aspect of students' HOTS. For example, the use of multiple-choice questions by many mathematics teachers. With this instrument, the teachers cannot identify which students have difficulties in transferring knowledge into new contexts, and in applying creative thinking and information literacy skills (Tanudjaya & Doorman, 2020). This will also complicate teachers to proceed on

[1]IAIN Syekh Nurjati, Jln. Perjuangan, Cirebon 45132, Indonesia, alimisri@syekhnurjati.ac.id

learning. In addition, the measurement inaccuracies will have an impact on students' self-efficacy. It is due that the students are considered unable to solve problems. Hence, the development of HOTS-based tests needs to fulfil the theoretical aspects of a test development, such as classical test and item responses theory.

Many researchers took parts in developing tests using the HOTS concept (e.g., Heong, et al., 2011; Mitana, Muwagga, & Ssempala, 2018; Mumu & Tanujaya, 2019; Rabadi & Salem, 2018; Tanujaya, 2016; Arifin & Retnawati, 2017; Bakry & Bakar, 2015). The developed HOTS-based tests are in the form of essay and multiple-choice questions, but they are for general mathematics skills. In their evaluation, some only use content validity and factorial analysis, while others only employ either classical test theory or item response theory. For instance, Tanujaya and colleagues developed the HOTS instrument applying expert validation, and factor analysis was carried out without using item analysis. Another research by Budiman and Jailani (2014) developed multiple-choice and essay HOTS tests with expert validation and classical test theory in determining the quality of the tests empirically. Hamdi, Suganda and Hayati (2018) also developed HOTS-based multiple-choice tests. The evaluation of the tests applied a content validity using Aiken’s formula and a reliability using Cronbach Alpha. The data of the empirical trials was analyzed using classical test theory, including the level of difficulty, items discrimination, and the functioning of distractors. Furthermore, Putri, Kartono, and Supriyadi (2020) developed a subjective test in the form of essay and analyzed the tests using item response theory with the Rasch model approach to evaluate the characteristics of the test and items. As for the aspect of thinking skills as the object under study, Bakry and Bakar (2015) highlight a different flow of thought processes. Mitana et al. (2018) measures thinking skills by dividing 3 levels of thinking processes, namely: remembering, understanding and arguing. The first two levels show low thinking skills (LOTS) while level 3 is for HOTS.

This study follows the HOTS level by Mitana et al. (2018), but the evaluation was simultaneously carried out using both classic and item response theories, which is considered different from the previous research. The items development procedure in this study refers to Beyers (2011). The test items, developed in the form of a subjective test (essay), are to measure students' HOTS on the matrix topic.

Theoretical Review

HOTS-based test instrument

In 1950, Bloom and his colleagues introduced a hierarchy of educational goals, known as Bloom's Taxonomy (Bloom et al., 1956). In their work, they did not specify the order or the complexity of thinking, only involving cognitive processes ranging from low-level of thinking skills to high-level ones. Several years later Anderson, Bloom's students and colleague revised the taxonomy. In the revision, it is shown that the highest level of thinking is not "evaluating" but "creating" (Anderson, et al., 2001; Krathwohl, 2002). Mitana et al. (2018) divides thinking levels into 3, namely: Level I (remembering), level II (understanding) and level III (arguing/reasoning). The first two levels show low thinking skills (LOTS) while level 3 is for high order thinking skills (HOTS). The ability to think at this level is to analysing, developing, and creating (Stanley & Moore, 2010).

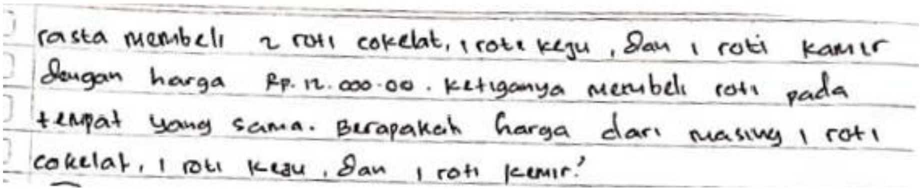
Tanujaya, Mumu and Margono (2017) use nine aspects of HOTS, namely: conceptual understanding, use of principles, impact prediction, problem solving, decision making, working within the limits of competence, facing/trying new things/challenges, having a pattern of thinking divergence, and lateral thinking patterns/imagination. The first five items measure critical thinking skills, and the last four items are for creative thinking. The instrument was developed using standard instrument development procedures, starting from developing conceptual definitions, operational definitions, determining constructs, dimensions, and indicators, to preparing blue prints, test items, expert validation, and testing. The research is based on critical and creative thinking processes. On the other hand, Bakry and Bakar (2015) divide thinking process into three levels: the ability to interpreting, making opinions and drawing conclusions. At the level of being able to interpreting, it is marked by the ability to collect information and write down the problem completely in questions. Meanwhile at the level of being able to make opinions, it is indicated by the ability to determine the required mathematical concepts and use them to solve problems. In addition, the level of being able to draw conclusions is characterized by the ability to determine the final answer and conclusion.

The HOTS indicator in the current study refers to Mitana et al. (2018), accounted for 19 aspects. The first two indicators are derived from Level I (remembering), the other nine indicators are derived from Level II (Understanding) and the remaining eight are derived from Level III (Arguing/reasoning). Level I consist of several operational indicators, including: repeating/imitating and recognising. Level II comprises two categories: understanding and applying. The understanding category is reduced to several operational indicators, including: interpreting, imitating, classifying, summarising, concluding, comparing, explaining. Meanwhile, the implementing category includes several operational indicators: executing/running, implementing/implementing. Level III consists of three categories, namely: analysing, evaluating and creating. From each category at level III, it is then reduced to operational indicators, including: analysing category, reduced to: distinguishing, organising, assigning attributes/marking. Meanwhile, the evaluating categories were reduced to: criticising and examining. Finally, the category of creating is reduced to: generating, planning and producing. Regarding the level of arguing/reasoning, according to Mumu and Tanujaya (2019), there are two categories: creative and imitation. Creative reasoning is divided into local and global. Meanwhile, imitation reasoning consists of rote reasoning (ordinary) and algorithmic (guided or limited).

The test items developed in this study were to measure students' HOTS on matrix topic and in the form of subjective test (essay-based questions). This is neither in Tanujaya et al. (2017) nor in Mitana et al. (2018) who developed the instrument in general, not on a particular subject, and it is in the form of multiple-choice questions.

HOTS is very important in learning. The effectiveness of each student’s learning depends on the learning process carried out while the learning process is very dependent on the her/his thinking ability. By having HOTS, students will be able to understand mathematical concepts in depth and apply them in real life (Bakry & Bakar, 2015). Students who have HOTS will be able to learn, improve performance and reduce their weaknesses (Heong et al., 2011). Furthermore, students who have HOTS are used to facing unusual problems, uncertainties, questions or dilemmas. That's why HOTS is believed to be able to better prepare students' quality of life in facing challenges both in advanced academic life and work and adult responsibilities every day (Rabadi & Salem, 2018). In this case, HOTS can be used to predict student success.

Content and pedagogic knowledge will make it easier for teachers to develop valid instruments in measuring students' mastery of material (Kristanto, Panuluh & Atmajati, 2020). Teachers' HOTS skills also determine how they transfer HOTS to their students (Misri, 2020). Analogical reasoning ability, one of the determining factors for HOTS (Richland & Begolli, 2016), can be used to develop HOTS instruments. These two aspects determine the thinking skills of the teacher and the way the teacher transfers it to students through the instruments they use. In fact, there are still teachers who still have difficulty designing the instrument (Budi & Junaini, 2018; Nurmasyitah & Hudiyatman, 2016). The following (Figure 1) is an example of a matrix test given by the teacher after learning. The test does not show any stimulant of knowledge transfer into a new context, in applying creative thinking.



rasta membeli 2 roti cokelat, 1 roti keju, dan 1 roti kamir
dengan harga Rp. 12.000.00. ketiganya membeli roti pada
tempat yang sama. Berapakah harga dari masing 1 roti
cokelat, 1 roti keju, dan 1 roti kemir?

Figure 1. A sample test given by a teacher after learning

Given the importance of measuring students' HOTS on a specific mathematics topic, it is necessary to develop matrix test items using the HOTS concept. The items formed must represent each topic of the matrix topic and meet the HOTS criteria (Anderson, et al., 2001; Mitana et al., 2018). The Items must also be able to distinguish which students have high HOTS, and master the content. In addition, the items must also have a variety of levels of difficulty (DeVellis, 2006).

Classical test and item response theory

A quality analysis of the test items developed in this study adapted a classical and modern test theory. The adaptation of these two theories has the same goal in developing a good test instrument by testing the abilities of prospective test takers (Partchev, 2004). The difference between the two in analyzing items can be seen from the results of the analysis. The results of the analysis using classical test theory only refer to the estimation of the difficulty index and the item discrimination index. The results of the analysis using modern test theory are more detailed, down to the item level. In classical test theory, the development of the test and its items is only based on the number of item response scores in the aggregate, while in modern test theory it is based on the specific characteristics of the item and is also based on the ability of each test taker (Thorpe et al., 2007).

Classical test theory is used to determine items discrimination, level of difficulty, validity and reliability (DeVellis, 2006; Magno, 2009). Meanwhile, modern test theory is used to determine the function of the characteristics of the items, especially the level of item difficulty and the ability of students to take tests (De Champlain, 2010). To complete the results of the analysis, the students' answers were also analyzed with reference to Bakry and Bakar (2015) to see their thought processes.

Methods

There are two major stages in this research: development of s HOTS-based test and items validation (Beyers, 2011; Bakker, 2019; Maharani, Sukestiyarno, Waluya & Mulyono, 2018). The first stage consists of: reviewing relevant literature, drafting test items, evaluating the draft by a team of experts and tryout. The second stage focuses on the validation of the items referring to the theory of classical test and item response.

Development of the test items draft

The draft step begins with interviews with high school students and a team of HOTS experts on the matrix topic related to their willingness and feasibility. This study involved 12 mathematics lecturers and 51 high school students. The lecturers are those who have special expertise related to the preparation of HOTS items for the matrix topic (expert team). Meanwhile, the selected students are those who have taken Mathematics lessons on the topic. Afterwards, a literature review was carried out regarding the HOTS instrument (Tanujaya, Mumu, & Margono, 2017; Bakry & Bakar, 2015; Mitana et al., 2018). The results of basic competences analysis on the topic are then used to construct items guideline (Manullang et al., 2017). The guideline is used as a reference for drafting the test items. For the items in the draft to be able to measure HOTS, the guideline is based on the HOTS criteria by Mitana et al. (2018).

To evaluate the drafted items, we prepared items review sheet, a scoring rubric, and answer keys. This data and the draft were then submitted to 12 mathematics lecturers, a team of experts, to get input so that the drafts made did not deviate. The expert's input is recorded on the item review sheet. The team was asked to check the suitability of the items and the guideline, in particular, based on the 19 HOTS criteria. The drafting ended with a tryout of the test items involving 51 high school students to obtain content validity. This test took about 90 minutes. Th students at this stage were given the HOTS test items on the matrix topic. Observations and data collection of student scores were carried out at this stage.

Validation of the test items

At this stage, the validity, reliability, level of difficulty and items discrimination were evaluated. The analysis was carried out based on the experimental results obtained using the classical test and item response theory. The approach used for item response theory is a Rasch Model. Analysis with classical test theory was carried out using the SPSS application, while analysis with the response test theory used the *winstep* application. The results obtained from the two approaches are used to determine the suitability of the matrix items and their feasibility in measuring HOTS.

Table 1. HOTS measurement criteria

Student's score	Category
$80 < \text{score} \leq 100$	Excellent
$60 < \text{score} \leq 80$	Good
$40 < \text{score} \leq 60$	Fair
$20 < \text{score} \leq 40$	Poor
$0 < \text{score} \leq 20$	Very poor

Prior to aforementioned tests - the analysis purpose, the assessment was carried out based on a scoring sheet that had been validated by a team of experts. The assessment is based on the ability to interpret, make opinions and draw conclusions (correct answers). Each test item is given the same scoring weight, which is 20. The total score obtained is on a scale of 1 - 100. The results of the calculation of the total score are interpreted using criteria in Table 1.

Table 2. Reliability criteria

r_α	**Interpretation**
0,80 – 1,00	Very high
0,60 – 0,80	High
0,40 – 0,60	Fair
0,20 – 0,40	Poor
0,00 – 0,20	Very poor

For classical test theory, a reliability is determined using the formula of *Cronbach's Alpha* and the criteria in Table 2 (Cautin & Lilienfeld, 2015).

$$r_\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma^2}\right) \qquad (1)$$

$\boldsymbol{r_\alpha}$ (Overall higher-order thinking skills test reliability); $\boldsymbol{\Sigma\sigma_i^2}$ (total variance score for each item); $\boldsymbol{\sigma}^2$ (total variance); $\boldsymbol{k}$ (number of items)

Next, the items discrimination index is calculated using the following formula and criteria (Table 3).

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} \qquad (2)$$

$\boldsymbol{D}$ (Items discrimination index); $\boldsymbol{B_A}$ (the number of participants in the higher-order thinking skills test who answered correctly in the upper group); $\boldsymbol{B_B}$ (the number of participants in the higher-order thinking skills test who answered correctly in the lower group); $\boldsymbol{J_A}$ (the number of participants in the upper group in the higher-order thinking skills test); $\boldsymbol{J_B}$ (the number of participants in the lower group in the higher-order thinking skills test)

The difficulty index is calculated based on the following formula and criteria (Table 4).

$$P = \frac{B}{JS} \qquad (3)$$

$\boldsymbol{P}$ (Difficulty index); $\boldsymbol{B}$ (the number of correct scores obtained by all students); $\boldsymbol{JS}$ (total overall score)

Table 3. The criteria of items discrimination

Interval value	Criteria
Negative – 0,09	Very poor
0,1 – 0,19	Poor
0,20 – 0,29	Fair
0,30 – 0,49	Good
0,5 – 1	Very good

Table 4. The criteria for difficulty index

Interval	Criteria
0 – 0,30	High
0,31 – 70	Medium
0,71 – 1,00	Low

Data analysis in this study also employed item response theory with the *Rasch model* approach. The theory uses a different approach from classical test theory in analyzing the items of a test instrument. According to De Champlain (2010), item response theory is a non-linear model that provides the probability of responding correctly to items as a function of item characteristics and the ability of test-takers. Meanwhile, according to Bond and Fox (2007), the Rasch model is a mathematical model that can measure the probabilistic relationship between the item's difficulty level and a person's ability.

The Rasch model can be used for data in format of dichotomy scores, scales, and the polytomy model. This study refers to the last data. The probabilistic function used for the polytomy model is as follows.

$$P_{ix}(\theta) = \frac{exp[\sum_{j=0}^{x}(\theta_n - \delta_{ij}]}{\sum_{r=0}^{m_i}\left[exp[\theta_n - \delta_{ij}]\right]} \qquad (4)$$

(Sumintono & Widhiarso, 2015, p.126)

$\boldsymbol{i}$ (Polytomy items with score category: $0,1,2,\ldots,\text{m}$); $\boldsymbol{\theta_n}$ (individual trait level (location of individual traits on the latent trait continuum**)**; $\boldsymbol{\delta_{ij}}$ (intersection of lines between categories (j) on items (i)); $\boldsymbol{P_{ix}(\theta)}$ (test-taker probability n score x with ability θ randomly selected can answer the item i correctly)

Findings and Discussion

This study resulted in 5 matrix test items based on the HOTS indicators; analysing, evaluating, and creating (Table 5). The five questions have been validated by 12 expert teams.

Table 5. Characteristics of items at level III

Level	Item	Total	Percentage
Analysing	1, 4, 5	3	60%
Evaluating	3	1	20%
Creating	2	1	20%

The items have also represented matrix topic indicators, as can be seen in Table 6.

Table 6. The sub-topics of matrix for the items

Items	Topics
1	Addition operations on matrix algebra
2	Subtraction operations on matrix algebra
3	Multiplication operations on matrix algebra
4	Inverse matrix
5	Determinant matrix

The data from the test results show all items are valid based on the value of ***r counting*** ≥ ***r table*** (see Table 7). This means that all the items made to measure HOTS for the matrix topic are appropriate.

Table 7. Summary of the validity for each item

Item	***r***		Validity
	r count	*r table*	
1	0,54	0,30	Valid
2	0,88	0,30	Valid
3	0,72	0,30	Valid
4	0,78	0,30	Valid
5	0,82	0,30	Valid

All of these items are reliable to measure students' mastery of the topic and HOTS, considering that the results are consistent if tried at different times (see detailed description in Table 8).

Table 8. Reliability score for each item

Item	Cronbach Alpha	Reliability
1	0,79	Reliable
2	0,67	Reliable
3	0,74	Reliable
4	0,79	Reliable
5	0,70	Reliable

Based on the results of the items discrimination test, it can be seen that each developed item is able to clearly distinguish between high and low-ability students, meaning that the test is able to distinguish which students have high HOTS or not.

Table 9. Summary of items discrimination for each item

Item	Test result	
	Index	**Category**
1	0,54	Good
2	0,88	Very good
3	0,72	Very good
4	0,78	Very good
5	0,82	Very good

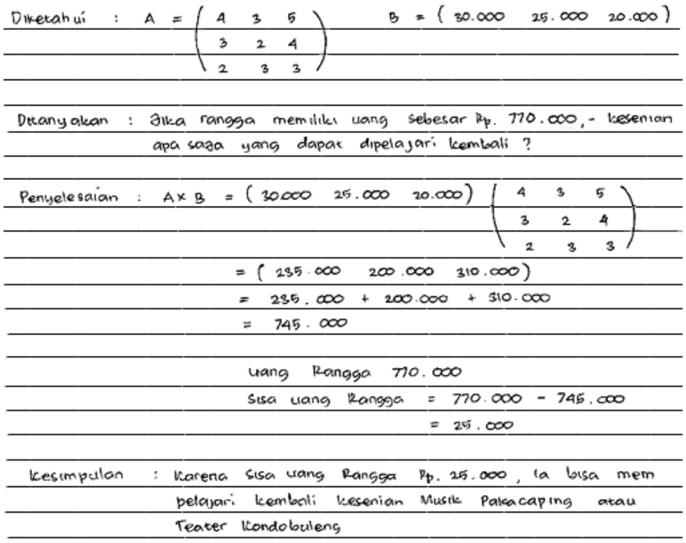
The difference in difficulty levels of the test items fall into low, medium, and high categories. The results of the analysis obtain 1 item with the low category (question number 2), 3 items with the medium category, (question number 1, 3, and 5), and 1 item with the high category (question number 4).

Table 10. The level of difficulty for each item

Item	Test result	
	Index	**Category**
1	0,69	Medium
2	0,71	Low
3	0,52	Medium
4	0,30	High
5	0,68	Medium

The findings indicate that the developed items have various levels of difficulty. With the differences, students will be grouped according to their ability levels.

The results of the items discrimination in Table 9 show a contrast on the students' answers. As an illustration, the answers to question number 2, as shown in Figure 2. The students' scores were quite varied; 20, 15, 10 and 5. From these findings, it can be seen that there is a significant difference between the highest scores and lowest ones, which means that the developed test is able to distinguish students with high and low abilities.



Diketahui : $A = \begin{pmatrix} 4 & 3 & 5 \\ 3 & 2 & 4 \\ 2 & 3 & 3 \end{pmatrix}$ $B = (30.000 \quad 25.000 \quad 20.000)$

Ditanyakan : Jika rangga memiliki uang sebesar Rp. 770.000,- kesenian apa saja yang dapat dipelajari kembali ?

Penyelesaian : $A \times B = (30000 \quad 25.000 \quad 20.000) \begin{pmatrix} 4 & 3 & 5 \\ 3 & 2 & 4 \\ 2 & 3 & 3 \end{pmatrix}$

$= (235.000 \quad 200.000 \quad 310.000)$

$= 235.000 + 200.000 + 310.000$

$= 745.000$

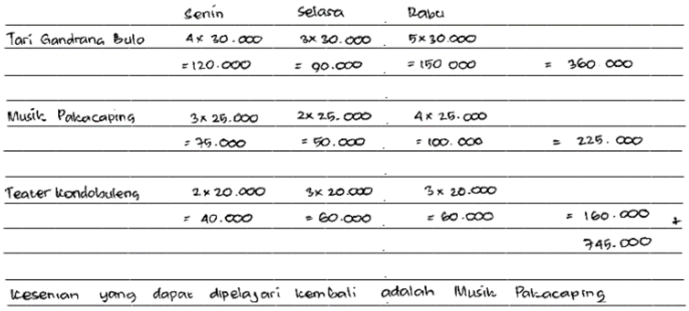
uang Rangga 770.000

Sisa uang Rangga = 770.000 − 745.000

= 25.000

Kesimpulan : Karena sisa uang Rangga Rp. 25.000, ia bisa mempelajari kembali kesenian Musik Pakacaping atau Teater Kondobuleng

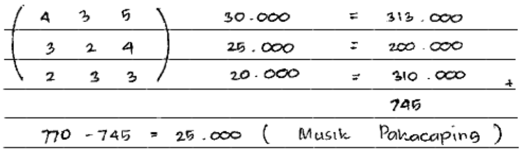
Figure 2a. A student's answer with score of 20



	Senin	Selasa	Rabu	
Tari Gandrana Bulo	4× 30.000	3× 30.000	5×30.000	
	=120.000	= 90.000	=150 000	= 360 000
Musik Pakacaping	3× 25.000	2× 25.000	4× 25.000	
	= 75.000	= 50.000	= 100. 000	= 225. 000
Teater kondobuleng	2× 20.000	3× 20.000	3× 20.000	
	= 40.000	= 60.000	= 60.000	= 160.000 +
				745.000

kesenian yang dapat dipelajari kembali adalah Musik Pakacaping

Figure 2b. A student's answer with score of 15

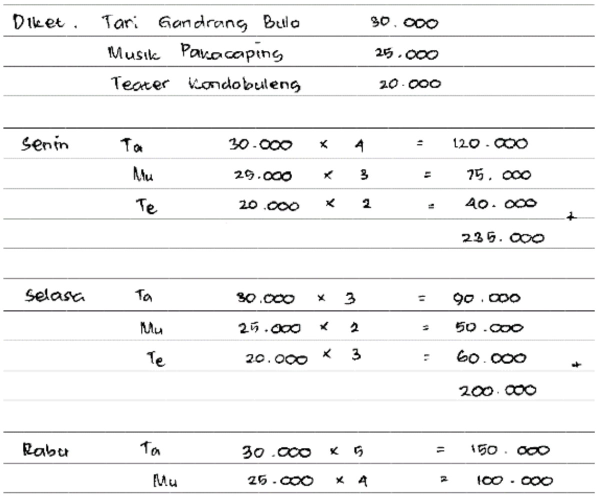


$$\begin{pmatrix}4 & 3 & 5\\ 3 & 2 & 4\\ 2 & 3 & 3\end{pmatrix} \quad \begin{matrix}30.000 & = & 313.000\\ 25.000 & = & 200.000\\ 20.000 & = & 310.000\end{matrix} +$$

745

770 - 745 = 25.000 (Musik Pakacaping)

Figure 2c. A student's answer with score of 10



Diket.	Tari Gandrang Bulo	30.000				
	Musik Pakacaping	25.000				
	Teater Kondobuleng	20.000				
Senin	Ta	30.000	×	4	=	120.000
	Mu	25.000	×	3	=	75. 000
	Te	20.000	×	2	=	40. 000 +
						235. 000
Selasa	Ta	30.000	×	3	=	90.000
	Mu	25.000	×	2	=	50.000
	Te	20.000	×	3	=	60.000 +
						200.000
Rabu	Ta	30.000	×	5	=	150. 000
	Mu	25.000	×	4	=	100.000

Figure 2d. A student's answer with score of 5

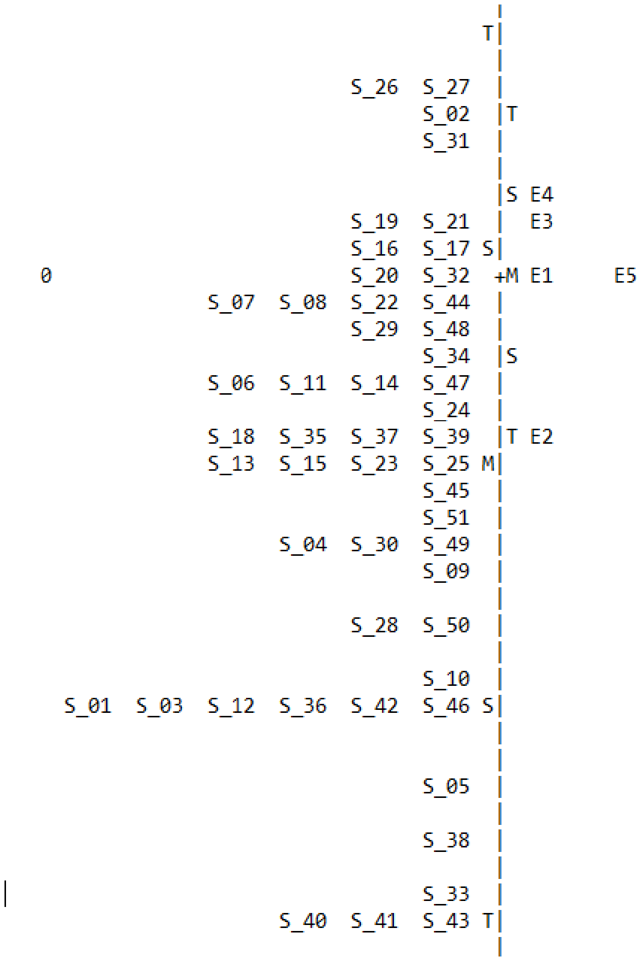
The student's answer in Figure 2d shows the incapability to interpret the test because the student has not been able to collect information and write the problem completely. This has an impact on the ability to make opinions and conclusions. In other words, the students have not been able to determine and use the necessary mathematical concepts so they cannot make decisions. Whilst the students' answers in Figure 2c seem to be able to collect information and determine the required mathematical concepts, though still insufficient. This ability has an impact on taking inappropriate decisions. It can be noted that the students in both figures, 2d and 2c, are in the category of poor ability.

In contrast, the students' answers in Figures 2a and 2b seem to be able to gather information and write down problems. They are also able to determine and use the necessary mathematical concepts to solve the problems. More specifically, the student in Figure 2a was able to solve the problem correctly and completely, while the student in Figure 2b was correct, but incomplete. These students are in high ability category.

The developed test of several items arranged with a level of difficulty that varies proportionally; low, medium, and high level of difficulty (Aiken, 2004). This means that a number of test items must contain items that are high, medium and low. This condition, the varying levels of difficulty, has met the criteria of the tests (further, see Table 10). The levels are able to describe all the student abilities. This is in line with Oermann and Gaberson (2016) who argue that the level of difficulty of each test depends on the ability of students to answer it.

Data analysis using item response theory

The distribution of students' abilities and the level of difficulty of the test items for measuring students’ HOTS can be seen in Figure 3.



```
                                         |
                                        T|
                                         |
                             S_26  S_27  |
                                   S_02  |T
                                   S_31  |
                                         |
                                         |S E4
                             S_19  S_21  |  E3
                             S_16  S_17 S|
   0                         S_20  S_32  +M E1      E5
                 S_07  S_08  S_22  S_44  |
                             S_29  S_48  |
                                   S_34  |S
                 S_06  S_11  S_14  S_47  |
                                   S_24  |
                 S_18  S_35  S_37  S_39  |T E2
                 S_13  S_15  S_23  S_25 M|
                                   S_45  |
                                   S_51  |
                       S_04  S_30  S_49  |
                                   S_09  |
                                         |
                             S_28  S_50  |
                                         |
                                   S_10  |
     S_01  S_03  S_12  S_36  S_42  S_46 S|
                                         |
                                         |
                                   S_05  |
                                         |
                                   S_38  |
                                         |
|                                  S_33  |
                       S_40  S_41  S_43 T|
                                         |
```

Figure 3. The distribution of students’ abilities and the difficulty levels of the items

Based on the Figure 3, it can be seen that most students have low abilities, and most of their abilities in working on questions are below the level of difficulty of the given test. This causes students to get low scores or think the items are very difficult. There are only four students whose abilities are higher than all levels of the given test, so that all four get the maximum score. Of the 51 test takers, 25 students had the ability to answer questions correctly, which was below the level of difficulty even for the items with the lowest difficulty level. Considering the level of difficulty of the items, there are three items spread across different levels of students’ ability, the remaining two items are at the same level of students’ ability, item 1 and 5. Overall, the difficulty level of the items is included in the good category because it is in the logit scale range 1 and -1 from the logit scale range 4 and -4 (Sumintono & Widhiarso, 2015). This means that the items in this range are not too difficult and not too easy if discussed independently without involving students' abilities. Also, the levels of difficulty are not that far apart. This is different from the ability of students who have a fairly wide range with a varying set of levels. Therefore, it can be concluded that the given test items can provide the information needed in the context of *assessment for learning.*

The test item's difficulty level can also be more precisely seen its value along with the level of item fit in Table 11 below.

Table 11. The analysis of items fit

Entry Number	Total Score	Total Count	Measure	Model S.E.	Infit		Outfit		PT-Measure		Exact Obs%	Match Exp%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	Corr	Exp			
4	422	51	0,11	0,04	1,69	2,9	1,54	2,4	0,72	0,69	15,7	14,3	E4
2	632	51	-0,18	0,03	1,02	0,2	1,22	1,1	0,79	0,70	3,9	11,2	E2
1	489	51	0,00	0,04	1,10	0,5	1,16	0,8	0,40	0,69	15,7	8,9	E1
5	487	51	0,01	0,04	0,73	-1,3	0,76	-1,2	0,71	0,69	9,8	8,2	E5
3	448	51	0,07	0,04	0,55	-2,5	0,54	-2,6	0,81	0,69	17,6	13,4	E3
MEAN	495,6	51,0	0,00	0,04	1,02	-0,1	1,05	0,1			12,5	11,2	
S.D	72,7	0,0	0,10	0,00	0,39	1,8	0,35	1,8			5,1	2,4	

Table 11 shows the level of difficulty of the items in order, from the highest to the lowest; 4, 3, 5, 1, and 2. The gap levels are not rather different. In addition, based on the PT-Measure, all items are positive. This shows that the items have the ability to distinguish students with high and low ability. However, in terms of the items suitability, item 4 (E4) contains one criterion that does not fit, the *infit mean-square* 1.69 which is greater than the value of 1.5. This means, the pattern of responses to the target items on test-taking students is less sensitive. In other words, the test-takers with certain abilities provide a pattern of answers to items that are not in accordance with their level of difficulty.

Furthermore, the level of ability and suitability of students' response patterns can be seen in Table 12.

Table 12. The table of Person statistics- misfit order

Entry Number	Total Score	Total Count	Measure	Model S.E.	Infit		Outfit		PT-Measure		Exact Obs%	Match Exp%	Person
					MNSQ	ZSTD	MNSQ	ZSTD	Corr	Exp			
36	30	5	-0,52	0,12	4,32	3,5	4,32	3,5	A -0,05	0,33	0,0	18,6	S_36
42	30	5	-0,52	0,12	4,32	3,5	4,32	3,5	B -0,05	0,33	0,0	18,6	S_42
41	15	5	-0,77	0,14	3,74	3,1	3,57	2,9	C -0,02	0,33	0,0	9,6	S_41
26	85	5	0,23	0,09	2,73	2,0	3,67	2,4	D -0,08	0,49	0,0	17,5	S_26
27	85	5	0,23	0,09	2,73	2,0	3,67	2,4	E -0,08	0,49	0,0	17,5	S_27
35	50	5	-0,21	0,12	3,24	2,4	3,32	2,4	F 0,08	0,34	0,0	13,8	S_35
39	50	5	-0,21	0,12	3,24	2,4	3,32	2,4	G 0,08	0,34	0,0	13,8	S_39
43	15	5	-0,77	0,14	2,39	2,0	1,98	1,5	H 0,93	0,33	0,0	9,6	S_43
38	20	5	-0,68	0,13	1,95	1,6	1,83	1,4	I 0,84	0,34	0,0	3,2	S_38

According to Table 12, it is identified that the student with the highest response had an ability score of 0.23 (student number 26 and 27), while the lowest with an ability score of -077 (students numbered 40, 41, and 43). In addition, there are 9 students' ability responses that contain one criterion that is not fit, the infit mean-square value is greater than 1.5. This means that the nine students, numbered 26,27,35,36,38,39,41,42,43, have a unique or inconsistent response pattern based on the level of difficulty of the items. That is, between the levels of students' ability to answer questions inconsistently with the level of difficulty of the items. Overall instrument analysis for students' ability level parameters can be seen in Table 13.

Table 13. The summary of measured items and persons

SUMMARY OF 51 MEASURED Person

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	48,6	5,0	-024	0,12	1,02	-0,2	1,05	-0,2
S.D.	18,0	0,0	0,27	0,01	1,10	1,5	1,16	1,5
MAX.	85,0	5,0	0,23	0,14	4,32	3,5	4,32	3,5
MIN.	15,0	5,0	-0,77	0,09	0,07	-2,9	0,07	-2,9
Real	RMSE	0,15 True SD	0,22	Separation 1,53		Person	Reliability	0,70
Model	RMSE	0,12 True SD	0,24	Separation 1,94		Person	Reliability	0,79
S.E. of Person Mean = 0,04								

Person Raw Score-to-Measure Correlation = 1,00
Cronbach Alpha (kr-20) Person raw score "test" reliability = 0,73

SUMMARY OF 5 MEASURED Item

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	495,6	51,0	0,00	0,04	1,02	-0,1	1,05	0,1
S.D.	72,7	0,0	0,10	0,00	0,39	1,8	0,35	1,8
MAX.	632,0	51,0	0,11	0,04	1,69	2,9	1,54	2,4
MIN.	422,0	51,0	-0,18	0,03	0,55	-2,5	0,54	-2,6
Real	RMSE	0,04 TRUE SD	0,09	Separation 2,21		Item	Reliability	0,83
Model	RMSE	0,04 TRUE SD	0,09	Separation 2,43		Item	Reliability	0,85
S.E. of Item Mean = 0,05								

Based on Table 13, the *person measure* (-0.24) is smaller than the logit value of 0.0. This shows that the tendency of students' abilities is smaller than the level of difficulty of the items. The reliability score 0.70 and the *Cronbach's alpha* 0.73 indicate that the consistency of the answers from students is good, while the item reliability 0.85 indicates that the quality of the items in the developed test has also good criteria. The quality of the student responses and good items was also strengthened by the acquisition of INFIT MNSQ and OUTFIT MNSQ scores. This can be seen that the values of both are 1.02 and 1.05 respective and are close to the ideal value of 1.00. Likewise, the INFIT ZTSD and OUTFIT ZTSD values, respectively, of -0.2 and -0.2 (for students) and -0.1 and -0.1 (for items) are also close to the ideal value of 0.00. The separation value of 1.02, based on the calculation, obtained a grouping value of 2 which means that there are two groupings of students based on their ability level. Meanwhile, the separation value for items of 2.2, after the calculation, obtained a grouping value of 3 which means that there are three groupings of items based on the level of difficulty.

The result of the item analysis using *Rasch Model* approach indicates all 5 test items are acceptable. This can be seen from the test results of the parameters of the student's ability level and the level of student difficulty which include: distribution of students' response patterns and item difficulty levels, student ability levels and item difficulty, order and suitability of students' ability levels based on response patterns and order and suitability of item difficulty levels. Therefore, in general, the developed test items are suitable to measure students' HOTS on the matrix topic.

The developed HOTS test in this study is only on the topic of matrix. Initially, this study used classical test theory analysis, which was then supplemented by the use of item response theory to examine not only at the characteristics of the items, especially the level of item difficulty but also at the test takers' ability to respond to the developed items. This research can be further developed by comparing the item analysis using classical theory and item response theory or being developed from other specific mathematics topics. The pattern of developing test items like this can be a reference for (prospective) mathematics teachers to be able to develop HOTS tests. More theoretically, the development of HOTS tests in this study can enrich the treasury of HOTS test items that fulfil psychometric characteristics, both related to the topics discussed and patterns of development that has been carried out.

Conclusion

This study develops 5 HOTS test items for a matrix topic, all of which are valid and reliable representing all HOTS indicators (Mitana et al., 2018). Based on the analysis using classical test theory, each item is proven to be able to distinguish which students have HOTS. The items also have various levels of difficulties. This means that in a set of items, the levels of difficulty are low, medium, and high. In addition, the items are able to describe all the students' abilities. The results of item response analysis using Rasch model approach, also provide an illustration that the five items are acceptable. This can be seen from the results of the students' ability parameter and students' difficulty levels, which include: the distribution of students' responses and item difficulty levels, students' ability levels and items difficulty, sequence and suitability of students' ability levels based on responses patterns, and order and suitability of item difficulty levels. Therefore, the developed test items are generally suitable, and can be used to measure students' HOTS on the topic of matrix.

Acknowledgment

The authors thank anonymous reviewers for their constructive reviews. Also, we are grateful for the support from the faculty and the mathematics education department IAIN Syekh Nurjati.

References

Aiken, L. R. (2004). *Assessment of intellectual functioning.* US: Springer Science & Business Media.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Arifin, Z., & Retnawati, H. (2017). Pengembangan instrumen pengukur higher order thinking skills matematika siswa SMA kelas X. *PYTHAGORAS: Jurnal Pendidikan Matematika, 12*(1), 98-108. Doi: 10.21831/pg.v12i1.14058

Bakker, A. (2019). *Design research in education: A practical guide for early career researchers.* New York: Routledge.

Bakry & Bakar, M. N. (2015). The process of thinking among junior high school students in solving HOTS question. *International Journal of Evaluation and Research in Education (IJERE), 4*(3), 138-145. Doi: 10.11591/ijere.v4i3.4504

Beyers, J. (2011). Development and evaluation of an instrument to assess prospective teachers' dispositions with respect to mathematics. *International Journal of Business and Social Sciences*, 20-32.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the social sciences* (2nd ed.). New York: Imprint Psychology Press. Doi: 10.4324/9781410614575

Budi, R. R. S., & Junaini. (2018). Peningkatan kemampuan guru matematika dalam perancangan soal-soal berbasis HOTS pada pelaksanaan supak melalui teknik pendampingan pengawas sekolah secara berkala. *Pakar Pendidikan, 16*(1), 60-83.

Budiman, A., & Jailani, J. (2014). Pengembangan instrumen asesmen higher order thinking skill (HOTS) pada mata pelajaran matematika SMP. *Jurnal Riset Pendidikan Matematika, 1*(2), 139-151. Doi: 10.21831/jrpm.v1i2.2671

Cautin, R. L., & Lilienfeld, S. O. (2015). *The encyclopedia of clynical psychology*. Malden, MA: John Wiley and Sons, Inc.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education, 44*(1), 109-117. Doi: 10.1111/j.1365-2923.2009.03425.x

DeVellis, R. F. (2006). Classical test theory. *Medical Care: Official Journal of the Medical Care Section, American Public Health Association, 44*(11), 50-59. Doi: 10.1097/01.mlr.0000245426.10853.30

Hamdi, S., Suganda, I. A., & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics. *REiD (Research and Evaluation in Education), 4*(2), 126-135. Doi: 10.21831/reid.v4i2.22089

Heong, Y. M., Othman, W. B., Yunos, J. B., Kiong, T. T., Hassan, R. B., & Mohamad, M. M. (2011). The level of Marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity, 1*(2), 121-125.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practise, 41*(4), 212-218. Doi: 10.1207/s15430421tip4104_2

Kristanto, Y. D., Panuluh, A. H., & Atmajati, E. D. (2020). Development and validation of a test instrument to measure pre-service mathematics teachers' content knowledge and pedagogical content knowledge. *Journal of Physics: Conference Series, 1470*(1). Doi: 10.1088/1742-6596/1470/1/012008

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Maharani, H. R., Sukestiyarno, Y. L., Waluya, S. B., & Mulyono. (2018). Design of creative thinking test in geometry based on information processing taxonomy model. *Beta: Jurnal Tadris Matematika, 11*(2), 144–155. Doi: 10.20414/betajtm.v11i2.180

Manullang, S., S, A. K., Hutapea, T. A., Sinaga, L. P., Sinaga, B., S, M. M., & Sinambela, P. N. (2017). *Matematika SMA kelas XI*. Jakarta, Indonesia: Pusat Kurikulum dan Perbukuan, Balitbang, Kemendikbud.

Misri, M. A. (2020). Propositional proofing techniques application in algebraic structure research. *Eduma: Mathematics Education Learning and Teaching, 9*(1), 1-13. Doi: 10.24235/eduma.v9i1.5624

Mitana, J. M. V., Muwagga, A. M., & Ssempala, C. (2018). Assessment of higher order thinking skills: A case of Uganda primary leaving examinations. *African Educational Research Journal, 6*(4), 240-249. Doi:10.30918/AERJ.64.18.083

Mumu, J., & Tanujaya, B. (2019). Measure reasoning skill of mathematics students. *International Journal of Higher Education, 8*(6), 85-91. Doi: 10.5430/ijhe.v8n6p85

Nurmasyitah & Hudiyatman. (2016). Kendala guru dalam merumuskan instrumen penilaian pada pembelajaran IPS sesuai dengan ranah afektif. *Jurnal Pesona Dasar, 2*(4), 48-62.

Oermann, M. H., & Gaberson, K. B. (2016). *Evaluation and testing in nursing education*. Springer.

Partchev, I. (2004). *A visual guide to item response theory*. Retrieved from https://www.metheval.uni-jena.de/irt/VisualIRT.pdf

Putri, B. S. F., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (hots) at high school mathematics students using the rasch model. *Journal of Research and Educational Research Evaluation, 9*(2), 58-69. Doi: 10.15294/jere.v9i2.46133

Rabadi, W. M., & Salem, R. K. (2018). The level of high-order thinking and its relation to quality of life among students at Ajloun University College. *International Education Studies, 11*(6), 8-21. Doi: 10.5539/ies.v11n6p8

Richland, L. E., & Begolli, K. N. (2016). Analogy and higher order thinking: learning mathematics as an example. *Policy Insights from the Behavioral and Brain Sciences, 3*(2), 160-168. Doi: 10.1177/2372732216629795

Stanley, T., & Moore, B. (2010). *Critical thinking and formative assessments*. New York: Routledge. Doi:10.4324/9781315856261

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan.* Cimahi, Indonesia: Trim komunikata.

Tanudjaya, C. P., & Doorman, M. (2020). Examining higher order thinking in indonesian lower secondary mathematics classrooms. *Journal on Mathematics Education, 11*(2), 277-300. Doi: 10.22342/jme.11.2.11000.277-300

Tanujaya, B. (2016). Development of an instrument to measure higher order thinking skills in senior high school mathematics instruction. *Journal of Education and Practice*, *7*(21), 144-148.

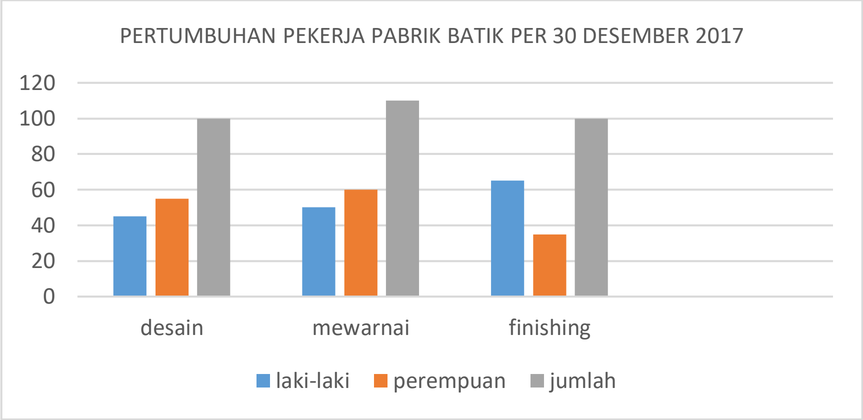
Tanujaya, B., Mumu, J., & Margono, G. (2017). The relationship between higher order thinking skills and academic performance of student in mathematics instruction. *International Education Studies*, *10*(11), 78-85. Doi: 10.5539/ies.v10n11p78

Thorpe, G. L., McMillan, E., Sigmon, S. T., Owings, L. R., Dawson, R., & Bouman, P. (2007). Latent trait modeling with the Common Beliefs Survey III: Using item response theory to evaluate an irrational beliefs inventory. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *25*(3), 175-189. Doi: 10.1007/s10942-006-0039-9

Zainudin, M., Subali, B., & Jailani. (2019). Construct validity of mathematical creativity instrument: First-order and second-order confirmatory factor analysis. *International Journal of Instruction*, *12*(3), 595-614. Doi: 10.29333/iji.2019.12336a

Appendix

Soal nomor 1



Dari tahun 2013-2016, jumlah pekerja laki-laki selalu lebih banyak dibandingkan jumlah pekerja perempuan pada salah satu pabrik batik lontara Makassar, Sulawesi Selatan. Pabrik tersebut membuka lowongan kerja hingga akhir desember tahun 2017 untuk penambahan jumlah pekerja pada bagian desain, mewarnai, dan *finishing* seperti terlihat pada grafik di atas. Jika lowongan diperpanjang hingga 30 Januari 2018 dan banyaknya pekerja perempuan yang mendaftar adalah 1 orang setiap kelipatan 30 jumlah pekerja yang ada, apakah jumlah pekerja laki-laki masih lebih banyak dari pekerja perempuan? Jelaskan alasanmu! Selesaikan dengan semua cara yang kamu bisa!

Soal nomor 2

Sulawesi Selatan, khususnya Makassar, telah membuka berbagai sanggar yang dapat diikuti oleh siapapun yang ingin mengembangkan bakat dalam kesenian. Sebagai pecinta kesenian, Rangga memutuskan akan melakukan pengembangan bakat kesenian khas Makassar selama 3 hari kedepan. Setiap kesenian yang diminatinya, sanggar menentukan harga Rp 30.000,- untuk kesian tari Gandrang Bulo, Rp 25.000,- untuk kesenian musik Pakacaping, dan Rp 20.000,- untuk kesenian teater Kondobuleng. Jika Rangga memiliki uang sebesar Rp 770.000,-, kesenian apa saja yang dapat dipelajari? Selesaikan dengan semua cara yang kamu bisa!

Tabel 4. Jumlah kesenian yang dipelajari Rangga selama 3 hari

	Senin	Selasa	Rabu
Tari Gandrang Bulo	4	3	5
Musik Pakacaping	3	2	4
Teater Kondobuleng	2	3	3

Tabel 5. Jumlah uang yang dikeluarkan Rangga

	Tari	Musik	Teater
Jumlah pengeluaran	30.000	25.000	20.000

Soal nomor 3

Rumah makan Pauh Piaman merupakan salah satu rumah makan yang menyediakan masakan tradisional di Sumatera Barat. Rumah makan tersebut sedang memberikan promo sebesar 30% untuk setiap 1 porsi makanan dan 50% untuk 1 gelas minuman. Siti dan teman-temannya memesan 3 gelas teh talua, 2 porsi soto Padang, dan 2 porsi pinyaram di rumah makan tersebut. Tak lama kemudian, Beni dan teman-temannya datang memesan 5 gelas teh talua, 1 porsi soto Padang, dan 3 porsi pinyaram. Terakhir, lala bersama teman-temannya datang memesan 2 gelas es teh talua, 2 porsi soto Padang, dan tidak memesan pinyaram. Siti harus membayar Rp 85.000, Beni Rp 80.000, sementara Lala Rp. 60.000 untuk semua pesanan mereka. Jika pembayaran Siti, Beni dan Lala belum termasuk promo yang diadakan rumah makan tersebut. Siapakah yang bisa menambah 3 porsi soto Padang dari sisa uang pembayaran? Berikan alasanmu!

Tabel 6. Jumlah pesanan

Makanan/ minuman	Siti	Beni	Lala
Teh talua	3	5	2
Soto Padang	2	1	3
Pinyaram	2	3	4
Pembayaran	85.000	60.000	80.000

Soal Nomor 4

Ibu Lia akan membuat 2 jenis kue tradisional Kalimantan Selatan, yaitu puracit banjar dan bingka barandam. Ia memiliki persediaan tepung 3.000 kg dan gula 2.000 kg. Bahan untuk membuat kue sudah disiapkan, yaitu: 3 kg tepung dan 2 kg gula. Kue puracit banjar memerlukan 150 gram tepung dan 50 gram gula, sedangkan kue bingka barandam memerlukan 100 gram tepung dan 100 gram gula. Modal awal ibu Lia Rp. 20.000 dan kue tersebut akan di jual oleh Bu Ani masing-masing seharga Rp 3.000. Dari pembagian hasil penjulannya antara Ibu Lia dan Ibu Ani sebesar 7 : 3, Ibu Lia mendapatkan keuntungan sebesar Rp. 32.000. Apakah pernyataan di atas benar? Jelaskan alasanmu! Jawablah dengan semua cara yang kamu bisa!

Soal Nomor 5

Hari jadi provinsi Jawa Timur diperingati setiap tanggal 12 Oktober yang bertepatan dengan hari ulang tahun Amira. Amira merupakan anak dari pak Andi yang menjabat sebagai walikota di salah satu kota di Jawa Timur. Pada tahun 2018, pak Andi didiagnosa Dokter hanya bertahan hidup 3 tahun lagi karena penyakitnya. Saat itu, umur pak Andi 28 tahun lebih tua dari umur Amira, umur bu Andi 6 tahun lebih muda dari pak Andi, sementara jumlah umur mereka bertiga 119 tahun. Apakah pak Andi masih bisa merayakan ulang tahun Amira ke-25 tahun yang bertepatan dengan hari jadi provinsi Jawa Timur? Jelaskan alasanmu! Kerjakan dengan semua cara yang kamu bisa!