Advances in Social Science, Education and Humanities Research, volume 545
Proceedings of the International Conference on Educational Assessment and Policy (ICEAP 2020)

Analysis of Difficulty Level and Discriminating Power Between Multiple Choices and Essay Items on Math Test

Rahmah Zulaiha*, Fahmi, Dian Rahdiani, Abdul Rahman, Muhammad Fathir Al Anfal
Centre for Assessment and learning, Research and Development and Books Agency
MOEC
Jakarta Pusat, Indonesia
*rahmahzulaiha@gmail.com, ffahmi6@gmail.com, dian.rahdiani@kemdikbud.go.id, rahman.abdul@kemdikbud.go.id, m.fathir239@gmail.com

***Abstract*—This analysis was intended to compare the level of difficulty and discriminating power on multiple choices and essay items on math test. The research was done by design treatments by subject's method. The research object was 22 items of multiple choices items and 22 essay items. The essay items were generated from multiple choices item that modified into essay items by eliminating the options. The essay items were divided into two clusters and each of them consist of 11 items. The population in this research was public and private Junior High Schools in DKI Jakarta Province, while the sample was defined by two-staged random sampling and stratified random sampling technique. There were 452 students act as the sample, taken from 6 private and public schools. Multiple choices item was scored by following rule; score 1 for correct respond and score 0 for incorrect respond, in meantime essay items were scored by using 3 different scoring rubrics, namely; Rubric Model 1 (0, 1), Rubric Model 2 (0, 1, 2) and Rubric Model 3 (key words). The difficulty level and the different power of students' response data that has been scored are calculated from the multiple choice and essay items, and then analysed by using descriptive statistical methods and inferential statistics and Rasch models. The results of the analysis with the Classic method obtained an average level of difficulty multiple choice questions greater than the essay items either on Model-1, Model-2, or Model-3. While the analysis using the Rash Model, the average level of difficulty multiple choice items is relatively the same as the essay items. For the average discriminating power, both using the Classic method and the Rasch Model, the average discriminating power of multiple choice items is greater than the essay items. In other words, based on research the form of multiple choice items is easier than the essay items, whereas the discrimating power of multiple choice items is greater than the essay items.**

Keywords—test, absorbency power, level of difficulty, discriminating power, rubric

I. INTRODUCTION

Education is the right of every individual as a citizen, and the government is obliged to facilitate this right. Education plays a very crucial role for the development of a nation. Nations that have superior and high qualified human resources are expected to compete with other nations in the world. In order to improve the quality of human resources, various efforts have been made by the Indonesian government, such as physical development in the form of facilities for the education implementation, as well as improving the competence of educators in each educational unit.

Schools as educational institutions play a very important role in developing abilities and transforming students into noble and knowledgeable human beings. The transformation of knowledge to students is inseparable from the role of educators in the teaching and learning process in the classroom. Learning is a key element in education, because by learning students can develop and increase their degree of life. According to Slameto, learning is a process of effort carried out by a person to obtain behavioural changes in the form of knowledge, skills, attitudes and positive values as an experience as a result of interaction with their environment [1]. Meanwhile, according to Marks learning is what students do, not what teachers do for students [2]. Thus it can be said that learning is an active process and aims to achieve a better life than before, not a passive process. So, the activities of the studying process are called learning.

According to Ahmad Susanto, learning is a teacher activity programmed in instructional design, to make students actively learn, which emphasizes the provision of learning resources [3]. In order for studying and learning activities to run well in accordance with the desired goals, educators must be equipped with an understanding of subject matter, learning theory, and assessment. Indicators of the education quality can be seen from the achievement of students in mastering the subject matter and this achievement can be seen through assessment. Assessment of student learning outcomes is carried out by educators to monitor the process, progress, and improvement of student learning outcomes continually.

According to the Ministry of Education and Culture [4], regarding the results of the 2015 PISA international study, Indonesian students' achievement in mathematics has increased significantly and gave quite optimism, although it is still lower than to the OECD average. In mathematics competence increased from 375 points in 2012 to 386 points in 2015. Despite the increase, Indonesian students still have difficulties and are not familiar to solve question that require reasoning. In line with the PISA study, Megawati et al [5] stated that from the results of their research, 73% of students in solving reasoning questions were in the poor category, especially the ability to evaluation questions.

Mathematics is a branch of science that plays a significant role in life. Not only mathematical concepts which are useful in problem solving, but also mathematical thinking patterns. Considering this important role, mathematics has become a compulsory subject starting from elementary school, it has even been introduced in kindergarten level. The goal of mathematics is not only to develop students' understanding on various mathematical concepts, but also to develop students' abilities in reasoning and problem solving. Through mathematics, students are expected to be able to think logically, systematically, critically, and creatively.

Students' absorption of subject matter that has been taught for a certain period of time can be assessed through tests. However, creating a good test instrument is not an easy work, it can be seen from the research results by Kartowagiran and Jaidun [6] and Hari Setiadi [7] which state that educators still do not understand and need guidance in assessment. In the PISA and TIMSS studies, various types of questions were applied, including multiple choices questions, short answers or filling in, complex multiple choice, and drag and drop. According to Umar and Hayat [8] the type of questions that are often used by educators in daily tests, 81.0% of them used essay questions and 56.5% of them still used multiple choice questions. Each type of questions has advantages and disadvantages. One of the disadvantages of the multiple choice questions is the possibility of guessing, while the disadvantages of essay questions is quite difficult to make the scoring rubrics.

One of scoring rubric function is to reduce the subjectivity of the scorer. With the scoring rubric, the scorer is expected to score consistently and omit subjectivity. There are several forms of scoring rubrics that can be used to score essay items, they possibly affect the scoring results including the analysis of items such as difficulty level and discriminating power. From the results of research conducted by Fahmi and Idris [9] on the use of three different scoring rubrics that are used to score essay questions, it can be concluded that there is no difference in the difficulty level of essay questions that are scored using different scoring rubrics. With the existence of several forms of scoring rubrics, it is necessary to carry out further research on the level of difficulty and discriminating power of essay questions and multiple choice questions on mathematics test.

Pujakusuma et al [10] found out that the main error in essay question is a conceptual error. The answers presented by students only show the problem solving procedure. No acknowledged and inquired statements and additional pictures. So that the answer seems cheating from his friend.

Among the students themselves, there was an inaccurate view, they consider essay questions are more difficult but when they were facing them, their learning methods were the same.

The essay question is a test (a set of questions in the form of assignments, questions) which requires students to organize and express the answers according to their own words (sentences). The answer can be in the form of recalling, arranging, organizing or combining the knowledge that has been learned in a well arranged series of sentences or words.

II. PROBLEMS

The problem raised in this study is whether there is a difference in the level of difficulty and the discriminating power between the multiple choice and the essay questions?

III. RESEARCH OBJECTIVES

Based on the above problems, the purpose of this study is to find out:

- Is there any difference on difficulty level between multiple choices and the essay questions that are scored with the scoring rubric model 1?
- Is there any difference on difficulty level between multiple choices and the essay questions that are scored with the scoring rubric model 2?
- Is there any difference on difficulty level between multiple choices and the essay questions that are scored with the scoring rubric model 3?
- Is there any difference on discriminating power between multiple choices and the essay questions that are scored with the scoring rubric model 1?
- Is there any difference on discriminating power between multiple choices and the essay questions that are scored with the scoring rubric model 2?
- Is there any difference on discriminating power between multiple choices and the essay questions that are scored with the scoring rubric model 3?

IV. THEORETICAL FRAMEWORK

Teacher often uses multiple choices and essay questions, for example in formative, subsumative or summative tests. On a larger scale, the use of these two type of question is used in international surveys such as TIMSS and PISA, while on a national scale multiple choices and essay questions are used in the Minimum Competency Assessment (AKM) as a substitute for the National Examination. According to Crocker and Algina [11], a test is a standard process for obtaining a sample of behaviour from a certain domain, while Umar [12] states that a test is a set of questions that must be answered, or

statements that must be chosen, responded to, or tasks that must be done by the person being tested (testee) in order to measure a certain aspect of behaviour of the testee. Saifuddin Azwar [13] said that a good test must be comprehensive and contain items that are relevant and comprehensive. Mehrens and Lehmann [14] classify the form of the question into two, namely the form of objective questions and non-objective questions (essay). Oosterhof defined multiple choice as questions that describe a problem and followed by a series of choices or alternatives [15]. Normally one choice is correct, while the other answer choices are distractors. The test taker's task is to choose one of the available answers, the correct one. According to Nitko [16] and Popham [17], multiple choices questions consist of two parts; the body (stem) and the alternatives of answer (option). Furthermore, Popham stated that the stem is divided into two, namely a direct question form and an incomplete statement form. Meanwhile, Mehrens and Lehmann [14] classify the essay question into a non-objective essay type and an objective essay type. The difference between objective essay type and non-objective essay lies in the certainty of scoring. According to Anas Sudijono [18], the essay item is a test in the form of a question or command that requires an answer in the form of a description or sentences which is generally quite long.

Meanwhile, Umar stated that the requirements for a good test include: (1) valid, means each measuring instrument only measures one dimension or aspect, and (2) reliability, means precision or accuracy of measurement results. According to Sugiono [19], reliability is a series of measurements or a series of measuring instruments that have consistency when they are applied repeatedly. Test reliability is the level of consistency of a test, how trustworthy a test can be to produce a consistent score, relatively unchanged even though it is tested in different situations. According to Best and Kahn [20], Sugiono [19], Sukadji [21] and Wiersma and Jurs [22] stated that the concept of reliability refers to the consistency of test score results in different conditions (place and time). According to Aiken and Lewis [23], if the test is to be used to determine the significance of the difference in the scores of the two groups of students, the reliability coefficient of 0.65 is considered satisfactory. If the test is to be used to compare students with one another, then at least a reliability coefficient of 0.85 is required. If the test is reliable, the information obtained from the test results can be trusted.

According to Wiersma and Jurs [24], the most important factor in assessing the accuracy of a learning achievement test is whether the items on the test are related to the knowledge that has been taught in the classroom. Meanwhile, according to Azwar [13], the quality of information obtained from the test results is determined by the quality of the test, and the quality of the test is determined by the quality of the items assembled in the test. Testing the quality of each item is carried out through item analysis, both qualitatively and quantitatively. Quantitative item analysis was carried out to determine the characteristics of the items, including the level of difficulty and discriminating power, as well as the test reliability. These item characteristics can be calculated by using classical and modern theory or the Rasch model. In the Rasch model the respond probability that answers an item correctly is modelled as a logistical function of the person and item parameters.

Crocker and Algina [11] said the difficulty level of an item refers to percentage of students who answered correctly. The more test takers who answered the item correctly, the easier the item was, in contrary the fewer students who answered the item correctly, the more difficult the item was. Each type of question will affect the level of difficulty. According to the results of research by Sucipto [25], the difficulty level of the short answer items is higher (more difficult) than the difficulty level of the multiple choice items, the discriminating power in the short answer items is higher than discriminating power in the multiple choice items. Meanwhile, according to Fahmi and Idris [9] regarding the essay items, there is no difference in the difficulty level of the essay questions that are scored using a different scoring rubric.

Scoring is the process of defining test results on a certain scale. Meanwhile, Nitko [16] states that there are two scoring methods used in essay questions, namely the analytic method of scoring based on keywords, values/ranges of values, or traits (the characteristics to be measured) and the holistic method (holistic), namely The scoring is done globally/generally, sorting (sequentially) or rating (scale).

In the Trends in International Mathematics and Science Study survey [26], the scoring of students' answers used a two-digit rubric. The first digit is used to determine the degree of students' correct answers and the second digit is used to classify the methods used by students in solving problems. The code used in scoring for the first digit is a maximum of 2, namely, if the answer is given without any errors (fully correct), a score of 1 is used for partially correct answers, and a score of 0 (zero) for wrong answers.

The difficulty level in the Rasch model and classical test theory is basically the same, the proportion of correct answers to the number of questions tested. The difference is that the opportunity value is then scaled by entering the logarithmic function. According to Hambleton et al [27], if ability is transformed so that the mean becomes 0 and the standard deviation is 1, then the difficulty level parameter is between -2.0 to +2.0. Items with difficulty level approaching -2.0 indicate easy questions, while questions with difficulty level approaching +2.0 indicate that the questions are difficult. The calculation of the discriminating power of the Rasch model questions or the correlation value of the item score and the Rasch score (*Pt Measure Corr*) is basically the same as the discriminating power of the classical test theory questions. In the classical test theory, the calculation uses a raw score and in *Pt Measure Corr* is used is a *measure* score. Theoretically, the discriminating power value lies between $-\infty$ and $+\infty$.

V. METHODOLOGY

This research is an experimental study using *design treatments by subjects*, an experiment that uses one group (one group experiment) that act both as experimental group and a

control group in different experiment periods. In this research, there was only one group of students as a control group who was subjected to a multiple choice and essay question type test as well as an experimental group when the essay question scoring was carried out using a different scoring rubrics.

The object of this research is 22 multiple choice items and 22 essay items derived from multiple choice items which their options were omitted. The essay items are divided into 2 clusters and each cluster consists of 11 essay items. The population in this study were all state and private junior high schools in DKI Jakarta, while the sampling was carried out by *two-stage random sampling and stratified random*, in the first stage, regional and sub-district random sampling was carried out using the drawing technique. After being drawn from 5 selected areas, the sample from South Jakarta City and from 10 Districts in South Jakarta, the sample from Pasar Minggu District was selected. The trial and research sample school were 6 public and private junior high schools selected from 24 junior high schools in Pasar Minggu District. The determination of the sample schools in this study was carried out by several steps. Firstly, all public and private junior high schools in Pasar Minggu District were grouped into 3 categories, namely with good, moderate, and poor quality junior high schools. From each group, 2 schools were randomly selected using a drawing technique. The trial sample class is class VIII and from each school 2 (two) whole classes are taken which are randomly selected by drawing technique.

There were 452 students participate as the sample. The data of students' result that had been scored were then used to calculate the level of difficulty and discriminating power of the multiple choices and essay items in mathematics test. The instrument was a mathematics test in multiple choice questions with four alternative choices and essay items.

The math essay items derived from a multiple choice test where the answer choices are omitted. The research data were analysed with the SPSS program, two-parameter Rasch Model, and Excel to determine the parameters of the difficulty level and the discriminating power. The data then analysed using descriptive statistical methods and inferential statistics. Descriptive statistics are used to describe, present, and inform the research data so that it is easy to read and understand, while inferential statistics are used to test the differences in the difficulty level parameters and the discriminating power of the two forms of tests, the multiple choice and essay.

VI. Results and Discussion

The following is an example of multiple choice questions and essay questions with three different scoring rubrics, namely Model-1, Model-2, and Model-3. The questions example is:

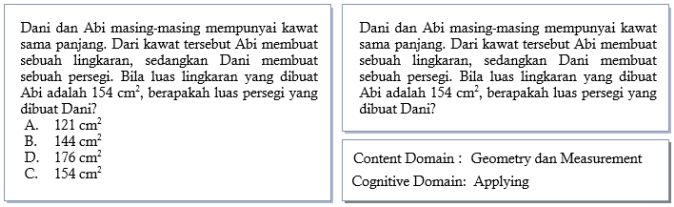

Fig. 1. Question example.

This figure 1 measures the student's ability to calculate the area of a wire square of a certain length, if it is known the area of the wire circle with the same length.

The maximum score for PG questions is 1, the maximum score for essay questions using the Model-1 scoring rubric is 1, Model 2 is 2, and Model-3 is 7. The following are the scoring rubrics Model-1, Model-2, and Model-3.

TABLE I. SCORING RUBRIC MODEL-1

No.	Criteria / Answer Key	Score
2	Luas lingkaran $= \frac{22}{7} \times r^2 = 154$ $r^2 = \frac{154 \times 7}{22} = 49$ $r = 7$ Keliling lingkaran $= 2 \times \frac{22}{7} \times 7 = 44$ Keliling persegi = keliling lingkaran = 44 cm Panjang sisi persegi $= \frac{44}{4} = 11$ cm Luas persegi $= 11^2 = 121$ cm^2	1
Maximum score		**1**

TABLE II. SCORING RUBRIC MODEL-2

No.	Criteria / Answer Key	Score
2	Jawaban benar 121 cm^2 beserta langkah penyelesaian Luas lingkaran $= \frac{22}{7} \times r^2 = 154$ cm^2 $r^2 = \frac{154 \times 7}{22} = 49$; $r = \sqrt{49} = 7$ cm Keliling lingkaran $= 2 \times \frac{22}{7} \times 7 = 44$ cm Keliling persegi = keliling lingkaran = 44 cm Panjang sisi persegi $= \frac{44}{4} = 11$ cm Luas persegi $= 11^2 = 121$ cm^2	2
	Ada indikasi perhitungan $\frac{22}{7} \times r^2$, $2 \times \frac{22}{7} \times r$ tetapi jawaban salah	1
	Jawaban benar 121 tanpa langkah penyelesaian, dengan atau tanpa satuan, jawaban dan langkah penyelesaian salah, kosong, ada coretan.	0
Maximum score		**2**

TABLE III. SCORING RUBRIC MODEL-3

No.	Criteria / Answer Key	Score
2	Luas lingkaran = $\frac{22}{7}$ x r^2 = 154	1
	$r^2 = \frac{154 \ x \ 7}{22} = 49$	1
	r = 7	1
	Keliling lingkaran = 2 x $\frac{22}{7}$ x 7 = 44	1
	Keliling persegi = keliling lingkaran = 44 cm	1
	Panjang sisi persegi = $\frac{44}{4}$ = 11 cm	1
	Luas persegi = 11^2 = 121 cm^2	1
Maximum score		**7**

Based on the scoring result, the percentage of students who answered correctly is shown in table 4.

TABLE IV. PERCENTAGE OF STUDENTS WHO ANSWERED CORRECTLY

Item No.	Percentage of Students who Answered Correctly			
	Multiple choice	***Model_1***	***Model-2***	***Model-3***
2	53,10%	19,19%	20,02%	20,89%

From table 4, it can be seen that the essay question that is scored with rubric Model-1, Model-2, and Model-3 categorized as difficult questions, while multiple choice question is categorized as medium questions. The mistakes that most students make in solving essay questions is misunderstanding the question (the length of the wire is the circumference) and incorrectly changed it into a mathematical model. In line with this research, according to the results of research conducted by Christina Khaidir and Elvia Rahmi [28] and research conducted by Aminah et al [29], the mistakes that most students make in solving story problems are carelessly read and understand the question, incorrect in making mathematical models and doing number operational. Meanwhile, according to Kusnita Damar Sari et al [30], 14,734% of students were able to solve math question require solving problems.

Following table shows the results of the analysis in the form of the average of difficulty level and the average discriminating power in the multiple choices and essay questions.

TABLE V. DIFFICULTY LEVEL AND THE DISCRIMINATING POWER IN MULTIPLE CHOICES AND ESSAY QUESTIONS

Question Characteristic	Method	Multiple choices	Essay		
			Model-1	***Model-2***	***Model-3***
Difficulty level	Classic	0,67	0,35	0,37	0,39
	Rasch Model	0,08	0,00	0,00	0,00
Discriminating power	Classic	0,51	0,37	0,37	0,37
	Rasch Model	0,49	0,37	0,37	0,36

From table 5, using the Classical Method, the average difficulty level of multiple choices questions is greater than the essay questions whether Model-1, Model-2, and Model-3. In other words, multiple choices questions are easier than essay questions. Meanwhile, using the Rash Model, the average difficulty level of multiple choices questions is relatively the same as the essay questions. For the average discriminating power between the questions using the Classical method and the Rash Model, multiple choices are greater than the essay questions.

A. Analysis Results with Classic Methods

The test results for the difference in the average difficulty level of multiple choices and essay questions scored with scoring rubrics Model-1, Model-2, and Model-3 are presented in table 6 below.

TABLE VI. TEST FOR DIFFERENCE IN THE AVERAGE DIFFICULTY LEVEL OF MULTIPLE CHOICES AND ESSAY QUESTIONS

Question type	Mean	Difference Mean	t	p
Multiple Choices	0,672564	0,3175773	50,292	0,000 ***
Essay Model-1	0,354986			
Multiple Choices	0,672564	0,3060091	40,963	0,000 ***
Essay Model-2	0,366555			
Multiple Choices	0,672564	0,2786000	40,586	0,000 ***
Essay Model-3	0,393964			

*p < 0,05, **p < 0,01, ***p < 0,001

Based on the test of the average difficulty level of multiple choices and essay questions scored using the scoring rubrics Model-1, Model-2, and Model-3 (table 6), it was obtained:

- There is a difference in the average difficulty level of the multiple choices questions (mean = 0.672564) and the essay questions using the scoring rubric Model-1 (mean = 0.354986) because p (0.000) <0.001.
- There is a difference in the average difficulty level of multiple choices questions (mean = 0.672564) and essay questions with the scoring rubric Model-2 (mean = 0.366555) because p (0.000) <0.001.
- There is a difference in the average difficulty level of multiple choices questions (mean = 0.672564) and essay questions with the scoring rubric Model-3 (mean = 0.393964) because p (0.000) <0.001.

The test results of the average discriminating power for multiple choices and essay questions scored with the scoring rubric Model-1, Model-2, and Model-3 are presented in table 7 below.

TABLE VII. TEST OF AVERAGE DIFFERENCE OF DISCRIMINATING POWER FOR MULTIPLE CHOICES AND ESSAY QUESTIONS

Question type	Mean	Differences Mean	t	p
Multiple Choices	0,505118	0,1338773	4,611	0,000 ***
Essay Model-1	0,371241			
Multiple Choices	0,505118	0,1355364	4,757	0,000 ***
Essay Model-2	0,369582			
Multiple Choices	0,505118	0,1348591	4,278	0,000 ***
Essay Model-3	0,370259			

*p < 0,05, **p < 0,01, ***p < 0,001

Based on the test for the average discriminating power in multiple choices and essay questions scored with the scoring rubric Model-1, Model-2, and Model-3 (table 7), it was obtained:

- There is a difference in the average discriminating power in the multiple choices questions (mean = 0.505118) and the essay questions with the scoring rubric Model-1 (mean = 0.371241) because p (0.000) <0.001.
- There is a difference in the average discriminating power in the multiple choices questions (mean = 0,505118) and the essay questions with the scoring rubric Model-2 (mean = 0,369582) because p (0,000) < 0,001.
- There is a difference in the average discriminating power in the multiple choices questions (mean = 0,505118) and the essay questions with the scoring rubric Model-2 Model-3 (mean = 0,370259) because p (0,000) < 0,001.

B. Analysis Results with Rash Model

The test results for the average difference in the difficulty level of multiple choices and essay questions scored with the Model-1, Model-2, and Model-3 scoring rubrics are presented in table 8 below.

TABLE VIII. TEST OF AVERAGE DIFFERENCE IN THE DIFFICULTY LEVEL OF MULTIPLE CHOICES AND ESSAY QUESTIONS

Question type	Mean	Difference Mean	t	p
Multiple Choices	0,0000	0,0000	0,0000	1,000
Essay Model-1	0,0000			
Multiple Choices	0,0000	0,00045	0,001	0,999
Essay Model-2	-0,0005			
Multiple Choices	0,0000	-0,00091	-0,002	0,998
Essay Model-3	0,0009			

*p < 0,05, **p < 0,01, ***p < 0,001

Based on the test for the difference in the average difficulty level of multiple choice and essay questions scored with the scoring rubric Model-1, Model-2, and Model-3 (table 8), it was obtained:

- There is no difference in average difficulty level of multiple choices questions (mean = 0,0000) and essay questions with scoring rubric Model-1 (mean = 0,0000) because p (1,000) > 0,05.
- There is no difference in average difficulty level of multiple choices questions (mean = 0,0000) and essay questions with scoring rubric Model-2 (mean = -0,0005) because p (0,999) > 0,05.
- There is no difference in average difficulty level of multiple choices questions (mean = 0,0000) and essay questions with scoring rubric Model-3 (mean = 0,0009) because p (0,998) > 0,05.

The test for the average discriminating power of multiple choices and essay questions scored with the scoring rubric Model-1, Model-2, and Model-3 3 presented in table 9 below.

TABLE IX. TEST OF AVERAGE DISCRIMINATING POWER OF MULTIPLE CHOICES AND ESSAY QUESTIONS

Question type	Question form	Mean	Difference Mean	t	p
Multiple Choices	PG	0,4936	0,12500	4,581	0,000 ***
Essay Model-1	Uraian Model-1	0,3686			
Multiple Choices	PG	0,4936	0,12682	4,752	0,000 ***
Essay Model-2	Uraian Model-2	0,3668			
Multiple Choices	PG	0,4936	0,13364	4,558	0,000 ***
Essay Model-3	Uraian Model-3	0,3600			

*p < 0,05, **p < 0,01, ***p < 0,001

Based on the test for the average discriminating power of multiple choices and essay questions scored with the scoring rubric Model-1, Model-2, and Model-3 (table 9), it was obtained:

- There is a difference in average discriminating power of multiple choices questions (mean = 0,4936) and essay questions with scoring rubric Model-1 (mean = 0,3686) because p (0,000) < 0,001.
- There is a difference in average discriminating power of multiple choices questions (mean = 0,4936) and essay questions with scoring rubric Model-2 (mean = 0,3668) because p (0,000) < 0,001.

There is a difference in average discriminating power of multiple choices questions (mean = 0,4936) and essay questions with scoring rubric Model-3 (mean = 0,3600) because p (0,000) < 0,001.

VII. CONCLUSION

The analysis results for the level of difficulty using the classic method show "There is a difference in the average

difficulty level of multiple choices and essay questions with the scoring rubric Model-1, Model-2, Model-3." Meanwhile, with the Rash Model show "There is no difference in the average difficulty level of multiple choices and essay questions with the scoring rubric Model-1, Model-2, Model-3." The analysis results for the discriminating power using the classic method and Rash Model show "There is a difference in average discriminating power of multiple choices and essay questions with scoring rubrics Model-1, Model-2, Model-3.

REFERENCES

[1] S. Slameto, Belajar dan Faktor-faktor yang Mempengaruhinya. Jakarta: PT. Rineka Cipta, 2010.

[2] J.L. Marks, A.A. Hiatt and E.M. Neufeld, Metode Pengajaran Matematika untuk Sekolah Dasar, terjemahan Bambang Sumantri. Jakarta: Erlangga, 1988.

[3] A. Susanto, Teori Belajar dan Pembelajaran di Sekolah Dasar. Jakarta; Kencana, 2013.

[4] Kementerian Pendidikan dan Kebudayaan, Peringkat dan Capaian PISA Indonesia Mengalami Peningkatan [Online]. Retrieved from: https://www.kemdikbud.go.id/main/blog/2016/12/peringkat-dan-capaian-pisa-indonesia-mengalami-peningkatan, 2016.

[5] H. Hartatiana, "Kemampuan Berpikir Tingkat Tinggi Siswa SMP dalam Menyelesaikan Soal Matematika Model PISA," Jurnal Pendidikan Matematika, vol. 14, no. 01, pp. 15-24, 2020.

[6] B. Kartowagiran and A. Jaidun, "Model Asesmen Autentik untuk Menilai Hasil Belajar Siswa Sekolah Menengah Pertama (SMP): Implementasi Asesmen Autentik di SMP," Jurnal Penelitian dan Evaluasi Pendidikan, vol. 20, no. 2, 2017.

[7] H. Setiadi, "Pelaksanaan Penilaian pada Kurikulum 2013," Jurnal Penelitian dan Evaluasi Pendidikan, vol. 20, no. 2, 2017.

[8] U. Umar and H. Hayat, Penelitian Penggunaan Soal Bentuk Obyektif dan Uraian di Sekolah. Jakarta: Pusat Penelitian dan Pengembangan Sistem Pengujian, 2000.

[9] F. Fahmi and H.M.N. Idris, "The analysis of students' error and difficulty level of mathematics essay test," International Journal of Educational Policy Research and Review, vol. 6, no. 4, 2019.

[10] G.K. Pujakusuma, "Analisis kesalahan pemecahan masalah matematika dan motivasi belajar siswa pada materi dimensi tiga," AKSIOMA: Jurnal Matematika dan Pendidikan Matematika, vol. 10, no. 2, pp. 172-179, 2019.

[11] L. Crocker and J. Algina, Introduction To Classical and Modern Theory. New York: Holt, Rinehart and Winston. Inc., 1986.

[12] J. Umar, Bahan Penataran Pengujian Pendidikan. Jakarta: Pusat Pengujian, 1998.

[13] A. Saifuddin, Tes Prestasi. Yogyakarta: Liberty, 1987.

[14] W.A. Mehrens and I.J. Lehman, Measurment and Evaluation In Educational and Psychology. New York: Holt, Rincchart and Winston, Inc, 1991.

[15] A. Ooesterhof, Classroom and Application of Educational Measurment, 2th ed. New York: Macmillan Canada Company, 1994.

[16] A.J. Nitko, Educational Assessment of Students, 2th ed. USA: Prentise-Hall, Inc, 1996.

[17] W.J. Popham, Modern Educational Measurement. New Jersey: Prentice-Hall, Inc, 1991.

[18] A. Sudijono, Pengantar Evaluasi Pendidikan. Jakarta: Raja Grafindo, 2001.

[19] S. Sugiono, Metode Penelitian Kualitatif. Bandung: Alfabeta, 2005.

[20] J. Best and J.V. Kahn, Research in Education. Boston: Allyn and Bacon, 1998.

[21] S. Sukadji, Menyusun dan Mengevaluasi Laporan Penelitian. Jakarta: UI-Press, 2000.

[22] W. Wiersma and S.G. Jurs, Research Methods in Education: An introduction. Boston: Allyn and Bacon, 2005.

[23] L.R. Aiken, Psychological Testing and Assessment. Massachusetts: Allyn and Bacon Inc, 1988.

[24] W. Wiersma and S.G. Jurs, Educational Measurement and Testing. Boston: Allyn and Bacon, 1990.

[25] S. Sucipto, Perbandingan antara Tipe Soal Jawaban Singkat dengan Pilihan Ganda Ditinjau dari Karakteristiknya (Tingkat Kesukaran, Daya Beda, dan Reliabilitas) pada Mata Pelajaran Bahasa Indonesia Tingkat SMP di Kota Bogor. Jakarta: PPs UNJ, 2006.

[26] OECD, TIMSS Field-test Scoring Guides Grade 8. Boston College, 2010.

[27] R.K. Hambleton, H. Swaminathan and H.J. Rogers, Fundamentals of item response theory. California: Sage Publications, Inc, 1991.

[28] C. Khaidir and E. Rahmi, "Analisis Kesalahan Siswa Dalam Menyelesaikan Soal Cerita Matematika Kelas X.2 SMAN 1 Salimpang Berdasarkan Metode Kesalahan Newman," Proceeding International Seminar on Education 2016 Faculty of Tarbiyah and Teacher Training, 2016.

[29] A. Aminah, R. Kiki and K. Ayu, "Analisis Kesulitan Siswa Dalam Menyelesaikan Soal Cerita Matematika Topik Pecahan Ditinjau Dari Gender," Jurnal Teori dan Aplikasi Matematika, vol. 2, no. 2, 2018.

[30] K.D. Sari, R. Rismayanti and I. Puspitasari, "Analisis kemampuan pemahaman dan pemecaham masalah matematik siswa mts pada materi bangun ruang sisi datar," JPMI (Jurnal Pembelajaran Matematika Inovatif), vol. 1, no. 5, pp. 965-974, 2018.