International Journal of Educational Policy Research and Review
Vol.6 (4), pp. 77-83 July, 2019
Available online at https://www.journalissues.org/IJEPRR/
https://doi.org/10.15739/IJEPRR.19.011

ISSN 2360-7076

Original Research Article

The analysis of students' error and difficulty level of mathematics essay test

Received 24 April, 2019 | Revised 19 June, 2019 | Accepted 24 June, 2019 | Published 21 July, 2019

Fahmi[1]
and
Idris HM Noor*[2]

[1]Examination Center, Office of Educational Research and Development, MOEC, Indonesia
[2]Research Centre, Office of Educational Research and Development, MOEC, Indonesia.

*Corresponding Author E-mail: hmnooridris7@gmail.com

Test is a way to know a students' learning achievement. The level of test difficulty plays a key role in determining the students' understanding of the learning material gained from the class. This study aimed at comparing the difficulty level of Mathematics essay test scoring from the rubrics test/a guide of different scores. This study analyzed 22 Mathematics essay test items at the Junior Secondary School (JSS). Design treatments by subjects uses one group experiment which means that there is only one group of students serving as the control and experiment group. A group of 452 third year students of JSS was given an essay test using different scoring or rubrics. Two-stage stratified random sampling was used to get the study sample. Scoring used three rubrics model 1 (0,1), model 2 (0,1,2), and model 3 (key words). The research revealed that the difficulty average of essay test scoring from the three rubrics/a scoring guide of model 3 is relatively similar. There is no significant difference between the level of essay test difficulty scored by rubrics/a scoring guide of models 1, 2 and 3. The most error of the students is to understand the items test, to replace the essay test into Mathematics model, and an error in operation of the variable.

Key words: Mathematics, essay test, difficulty level rubrick, a scoring guide, junior secondary school.

INTRODUCTION

Assessment from a teacher plays an important role in knowing students' learning achievement and the effectiveness of teaching and learning activities in class. The results of an assessment give accurate information on how far students have gained and understood the teaching and learning process. By understanding the progress of students' learning achievement, a teacher can give reinforcement or enrichment to students.

Umar and Hayat (2000) found out that the use of multiple choice and essay test by teachers in posttests in teaching and learning activities ranges 81% for the essay test, fulfilling test (63.9%), and multiple choice (56.5%).

The finding revealed that teachers tend to use essay test for both formative and sumative tests more often. The essay test is also used by the Trends in International Mathematics and Science Study (TIMSS) and the programme for International Students Assessment (PISA) to measure students' competence in Mathematics, language, and science and most students find it difficult to do the essay test in Mathematics (Wahyuddin, 2016, Sulestry and Meliyana, 2018). One of the weaknesses of essay test is the difficulty in setting up the scoring rubrics due to the different difficulty levels of the test items.

The question is " Is there any different difficulty level of scoring from three different rubrics/scoring guide?. This study analyzed 22 Mathematics essay test items at the Junior Secondary School (JSS) level. Therefore, it is important to understand students' error in Mathematics

essay test and to determine the difference of scoring difficulties of essay test from the different level of difficulties of the test items using rubrics /scoring guide.

Literature Review

One of the measuring instruments for obtaining the information of students' learning achievement is a test. An instrument is a systematic procedure for measuring the sampling behavior (Linn & Gronlund, 1990) and the quality of information is determined by a test (Azwar, 1987). The test is a process of obtaining the behavior of a certain sample (Crocker and Algina, 1986; Allen, 1979) using a systematic procedure to observe students behavior in a numeric scale (Cronbach, 1960), standard process to obtain the sample of behavior from certain aspects of a person's behaviour (Algina, 1986; Allen, 1979). On the other hand, Anastasi (1988) stated that a test is a result of learning achievement or performance to measure the student's knowledge as a result of educational program or training while the test quality is determined by qualitative and quantitative test items. A good test refers to validity and realibility as well as being objective, non discriminative, comprehensive, and easy to use (Suryabrata, 1987). Teachers often utilize essay test in formative and summative evaluation, but the large scale survey of students' ability has been conducted by Trends in International Mathematics and Science Study (TIMSS) and The Programme for International Students Assessment (PISA) (OECD, 2010).

A test as a standard tool should fulfill the condition of obtaining information of a person's behavior or to measure the length of things such as ruler and the weigher or heavy equipment to measure the weight. The test condition should be reliable and valid which means carefully and accurately done (Umar, et.al, 1997) and the important factor in assessing the student's achievement is whether the test items measure the knowledge taught in the class (Wiersma and Jurs, 1990). A good test should be comprehensive and the test items should contain reperesentatively and proportionally all materials taught in the class (Azwar, 1987).

Mehrens and Lehmann (1991) classified the test item into objective and non objective essay test. The difference between these tests is that the objective test is the question to get a certain correct or incorrect answer while the non-objective test is more challenging to give some possible answers or the answer is the description of an idea of the testee (Zainul and Noehi, 1997). For the scoring, the objective test is more accurate while non objective test is subjective which means the scoring sometimes depends on the view of the tester or evaluator. To minimize the subjectivity of scoring, it has to be made a clear detailed guide of scoring which makes the scoring of the two relatively similar. It has also been suggested that the test item should not be too difficult or too easy. The too difficult test item cannot express what the learners know about the question while the too easy test item cannot express what is unknown to the learners (Joni, 1986), and the level of difficulty of the test items is the percentage of the student's correct answer (Allen, 1979; Wiersma and Jurs, 1990; Crocker and Algina, 1986). The greater the number of students that answer the test correctly, the easier the test items while the fewer the number of students that answer the test items, the more difficult the test items. It means that the level of test items is more related to the level of test difficulty.

This study uses the formula of proportion to determine the level of difficulties of essay test items. The first step to determine the level of difficulty of a test item is to score every test item based on the score guide or scoring rubrics. Scoring is a process of changing an answer of the instrument into number which is a quantitative value from an answer of an item in instrument (Djaali and Muljono, 2008), while Suryabrata (1987) explained that the scoring is a process of definition of the test result into a certain scale. There are two methods used in essay test, an analytical method is a method of scoring based on the key words, value/range value or trait and holistic method is the global scoring, sorting or rating. The Ministry of Education and Culture (1993) defines a scoring guide as a guide containing the possibilities of correct answers or key words and the value of scoring determined for each key answer. Next, in scoring, an objective essay is only possible for two categories (right or wrong). For each key, the right word is scored 1 (one) and the wrong one is scored 0 (zero). One formula of the answer can contain more than one key word so that the maximum score of one answer is 1 (one) or more. Each unanswered key word or wrong answer is scored 0 (zero). Scoring of the student's answer uses two digit rubrics (OECD, 2010). The first digit is used to determine the degree of truth of the students' answer, and the second digit is used to classify methods used by students in overcoming their problems. A code used to score for the first two digit is maximum 2 (two) if the answer is without any errors (full value), score 1 (one) is for a partly correct answer and 0 (zero) score is for a wrong answer.

This study employed three rubrics/scoring guides, rubrics/scoring guide model 1 (0.1) is that every formula of student's correct answer is given score 1 (0.1) while every students' correct answer is scored 1 (one) and students' wrong answer is scored 0 (zero). The rubrics/scoring guide model 2 (0,1,2), the score is maximum 2 if all answers are correct, the score is used for some correct answers, and 0 (zero) for wrong answer or zero answer. Rubrics/scoring guide model 3 (key word), every key word of the student's answer is to give a score.

METHODOLOGY

This study used design treatments by subjects of one group experiment for both the experiment and control groups in other different periods of experiment. This experiment only used one group of students as a control group and were

Table 1. The average of the difficulty level of essay test based on the scoring rubrics

Rubrics/Scoring Guide	Level of Difficulty
Model 1	0515
Model 2	0.3656
Model 3	0.3940

Example 1:

A block framework made from wire in similar footage $\frac{5}{4}$ wide and tall similar with $\frac{3}{4}$ wide. The length of the wire used to make the framework of block is 240 cm. How much is the volume of its block frame?.

Content Domain: Geometry

Deskripion : to calculate volume of block frame

Rubrics/scoring guide Model 1

No. Test item	Criteria/Answer key	Score
	Width (ℓ) = y, length (p) = $\frac{5}{4}$ y, height (t) = $\frac{3}{4}$ y Length of wire = 4(p + ℓ + t) = 240 = 4($\frac{5}{4}$ y + y + $\frac{3}{4}$ y) = 4 $(\frac{5+4+3}{4})$ y 12y = 240 y = 20 Length = $\frac{5}{4}$ x 20 = 25 cm, tinggi = $\frac{3}{4}$ x 20 = 15 cm Block volume = 25 x 20 x 15 = 7.500 cm^3	1
	Maximum Score	1

given an essay test and as an experiment group they were scored using different scoring rubrics. The object of this study is 22 essay item tests divided into two clusters, each of them consists of eleven descriptive test items. The population is all schools (JSS) in selected areas of the city. Sampling was done by two-stage random sampling and stratified random sampling. The sample consists of six public and private JSS. Each school had two eight grade classes. The data of students who have been scored was calculated based on the difficulty level of each essay item test. Then, it was analyzed using descriptive and inferential statistics.

RESULT AND DISCUSSION

The result of item test analysis from the three rubrics/a scoring guide, the average of the difficulty level can be seen in the Table 1.

Table 1 shows that the average difficulty level of the essay test which is scored from three rubrics/a scoring guide is relatively similar. An example of the questions and the level of difficulty of the test items (a percentage of correct answers) from the scoring of the three rubrics, is as follows:

The item test is to measure the ability of the students in measuring block frame which consists of length, width, and height in the form of variables.

Based on the result of scoring using rubrics/scoring guide model 1, the average of students who can answer correctly is 19.19%. Specific test item including the category of difficult test items with 0.199 level of difficulty. It means that students still find it difficult to do the algebra test. The error of what the students do in algebra consists of the errors in negative sign and the error in the form of algebra. Rubrics/Scoring guide Model 2.

Based on the scoring of the rubrics/scoring guide model 2, the average of students who answer correctly about the

Rubrics/scoring guide Model 2

	2	Correct answer 7.500 cm^3 and the step of doing it Width (ℓ) = y, length (p) = $\frac{5}{4}$ y, height (t) = $\frac{3}{4}$ y The length of wire = 4(p + ℓ + t) = 240
		= 4($\frac{5}{4}$ y + y + $\frac{3}{4}$ y)
		= 4 $(\frac{5+4+3}{4})$ y = 12y = 240 Width (y) = 20 cm Length = $\frac{5}{4}$ x 20 = 25 cm, height = $\frac{3}{4}$ x 20 = 15 cm Block volume = 25 x 20 x 15 = 7.500 cm^3
	1	The indication of calculation 4(p + ℓ + t) = 4($\frac{5}{4}$ y + y + $\frac{3}{4}$ y) = 240 There is an error in determining the width (y), but it is consistent to use value Y in the following calculation Length = $\frac{5}{4}$ x .... and Height = $\frac{3}{4}$ x ....
	0	Correct answer 7.500 without a step to finish, answer and the steps of incorrect finishing, zero, and without finishing, the step error finishing, and zero.

Rubrics/Scoring guide Model 3

No. Question	Criteria/Key answer	Score
	Width (ℓ) = y, length (p) = $\frac{5}{4}$ y, height (t) = $\frac{3}{4}$ y	1
	The length of wire = 4(p + ℓ + t) = 240	1
	= 4($\frac{5}{4}$ y + y + $\frac{3}{4}$ y)	1
	= 4 $(\frac{5+4+3}{4})$ y	1
	= 12y = 240 y = 20	1
	Length = $\frac{5}{4}$ x 20 = 25 cm, tinggi = $\frac{3}{4}$ x 20 = 15 cm	1
	Block volume = 25 x 20 x 15 = 7.500 cm^3	1
	Maximun Score	7

test items is 20.02%. The test item includes the difficult test item at the 0.2002 level of difficulty.

Based on the scoring done by rubrics/scoring guide model 3, the average number of students who answered the questions correctly is 20.89%. This includes difficult questions at the 0.2089 level of difficulty. The most common mistake of the students in doing the test is the mismatch of understanding the questions and error in changing into Mathematics model.

The most common mistake done by students in doing the test is that the students understand wrongly and make mistakes in changing the question into the Mathematics model form. The mistake made by students when doing essay test is that they are not careful to read and understand the questions, thereby making error in making the Mathematics model, and error in doing cardinal and

Table 2: Test the difference of two averages of the difficulty level from the scoring rubrics models 1 and 2, as follows.

Group Statistics					
	Model	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Difficulty level	1	22	.354986	.2240613	.0477700
	2	22	.366555	.2337150	.0498282

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference**	
									Lower	**Upper**
Difficulty level	Equal variances assumed	.126	.725	-.168	42	.868	-.0115682	.0690277	-.1508718	.1277354
	Equal variances not assumed			-.168	41.925	.868	-.0115682	.0690277	-.1508791	.1277427

Table 3. Test the difference of two average difficulty levels of test essay based on the scoring rubrics Models 1 and 3.

Group Statistics					
	Model	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Difficulty level	1	22	.354986	.2240613	.0477700
	3	22	.393964	.2283923	.0486934

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Difficulty level	Equal variances assumed	.010	.921	-.571	42	.571	-.0389773	.0682131	-.1766368	.0986823
	Equal variances not assumed			-.571	41.985	.571	-.0389773	.0682131	-.1766383	.0986838

ordinal numbers (Rahmi, 2016; Aminah et al., 2018), lack understanding of Mathematics essay test and in the construction of a Mathematics model (Maulana and Hasnawati, 2016), limited time in doing the test, concessive, inaccurate, often forget, feel anxious, in a hurry to do the test (Novferma, 2017). However, if the students are trained continuously to solve the problem, they will enhance their Mathematics literacy (Fajaruddin et al., 2018). The following group of statistics explains the difficulty level models 1 and 2.

Table 2 shows, if sig. > α, so it can be concluded that there is no difference in the difficulty level of the essay test scored by rubrics/scoring models 1 and 2. From the result of the data analysis of the test, the difference of two average difficulty levels from scoring rubrics model 1 and model 3, can be seen in Table 3.

Table 3 shows, if sig. > α, so it can be concluded that there is no difference in the difficulty level of the essay test scored by rubrics/scoring models 1 and 3. From the result of the data analysis of test the difference of two average difficulty levels from scoring rubrics models 2 and model 3, are shown in Table 4.

Table 3 shows, if sig. > α, so it can be concluded that there is no difference difficulty level of essay test scored by rubrics/scoring models 2 and 3.

Conclusion

Based on the level of difficulty of the essay test, it shows that the average difficulty level of the essay test with scoring guide 1= 0.3515, the average difficulty levels of the essay test based on scoring model 2 = 0.3656, and the average difficulty levels of the essay test based on scoring model 3 = 0.3940. It means that the average level of difficulty of the essay scored from the three models of the scoring guide is relatively similar. There is no significant difference between the level of difficulty of essay test scored by rubrics/scoring guide model 3. The kind of error

Table 4. Test the difference of two average difficulty levels of test essay based on the scoring rubrics Models 2 and 3.

Group Statistics					
	Model	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Difficulty level	2	22	.366555	.2337150	.0498282
	3	22	.393964	.2283923	.0486934

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Difficulty level	Equal variances assumed	.062	.804	-.393	42	.696	-.0274091	.0696699	-.1680087	.1131905
	Equal variances not assumed			-.393	41.978	.696	-.0274091	.0696699	-.1680109	.1131927

involving the most students in essay test includes understanding the question, transferring the essay questions into a Mathematics model, and the error in operation in the form of a variable. For a routine test item such as calculating the prisma volume and volume of a cube, more than 70% had the correct answer.

Triangular base area = $\frac{1}{2}$ x 12 x 9 = 54 cm^2

Prisma volume = base area x high = 54 x 10 = 540 cm^3

Most students can answer the question using one or two simple techniques.

Acknowledgement

The authors are grateful to the Office of Educational Research and Development, Ministry of Education and Culture (MOEC) for encouraging the researchers to conduct the study. The authors are also thankful to the Head of Examination Centre, Office of Educational Research and Development who promoted and supported the researchers to attend the workshop delivering this research at other institutions within the MOEC. The authors are very grateful to the resource persons at the provincial and district educational levels and teachers who have given the data and information of the test.

Author contributions

Fahmi and Idris HM Noor designed the research, performed the research, and analyzed the data. Both authors wrote the paper, proofread it, and approved the final manuscript.

Conflict of interests

The authors declare that they have no conflict of interests.

REFERENCES

Aminah A, Kiki R, Ayu K (2018). An analyisis of students' problem in doing the mathematics essay test: fraction reviewed from gender. J. Mathema. Theory Application. 2 (2):118-122..

Anastasi A(1988). Psychological Testing, 6th ed, New York: Macmillan Publishing Company.

Azwar S (1987). Achievement Test. Yogyakarta: Liberty.

Crosker LM, James A (1986). Introduction To Classical & Modern Test Theory. Holt, Rinehart dan Winston. Inc. New York.

Djaali, Muljono P (2008). Measurement in Education, Jakarta: Post Graduate Public Jakarta University.

Fajaruddin MA, Rahmita YG, Nareki ML (2018). The influence of Problem Solving Approach towards the Students' Ability of Mathematics Literacy 5(2):135-146.

Gronlund NE, Linn DRL, (1990). Measurement and assessment in teaching. New York: Macmillan Publishing Company.

Joni TR (1986). Measurament and Educational Evaluation. Surabaya: Karya Anda.

Khaidir C, Rahmi E (2016): Students Error Analysis in doing Mathematics Essay Test of X.2 Garde Public Senior Secondary School 1 Salimpang, on Newman Error Analysis. Proceeding at International Seminar on Education 2016, Faculty of Tarbiyah and Teacher Training. campus.iainbatusangkar.ac.id/ ojs/index.php/ proceedings/ article/ download/ 630/622

Lee J. Cronbach (1960). Essential of Psychological Testing, 3th ed, New York: Harper & Row, 1960.

Malihatuddarojah D, Rully C, Indra P (2019). An Analysis of Students' Error in Operational Form of Alzebra. Journal of Education. 13 (1):1-8.

Mary J Allen (1979). Introduction Measurment Theory. California: Montrey.

Maulana A, Hasnawati. (2016). A Description of Mathematics Literary Ability of VIII Grade Public Junior Secondary School Students (JSS) Kendari. Res. J. Mathematics Educ. 4 (2): 1-14

Mehrens WA, Lehman IJ (1991). Measurment and Evaluation In Educational and Psychology. New York: Holt, Rincchart and Winston, Inc.

Ministry of Education and Culture, (MOEC) Indonesia. (1993). Drafting, Scoring, and the Use of Achievement Test Learning Form Description. Jakarta: Examination Centre.

Nitko, Anthony J (1996). Educational Assessment of Students. Engliwood Cliffs, Prentise-Hall, Inc.

Novferma N (2017). Difficulty Analysis of Junior Secondary School Students' Self-efficacy in to Solve Essay Mathematics Problems. Res. J. Mathematics Educ. 3 (1):76-87

OECD. (2010). TIMSS Field-test Scoring Guides Grade 8. Boston Collge.

Sulestry AI, Meliyana SM (2018). An Analysis of Ability to completeMathematics Essay Test on VII grade Public Junior Secondary School (JSS) Students I Bulukumba. A Proceeding of National Seminar. 03 (1): 212-220.

Suryabrata S (1987). The Development of Learning Outcomes Test. Jakarta: Rajawali.

Umar J, Hayat B (2000). A Research on the Use of Objective Test and Essay Test in School. Jakarta:Examination Centre, Office of Educational Research and Development, Ministry of Education and Culture.

Umar J, Haribowo H, Hayat B, Akhmad AM (1997). A Training Material of Educational Evaluation, Jakarta: Examination Centre.

Wahyuddin (2016). An Analyisis of the Ability in Doing Mathematics Essay Test Based on the Verbal Ability. 9 (2):148-160,.

Wiersma, William and Stephen GJ (1990). Educational Measurment and Testing, 2nd ed. Boston: Allyn and Bacon.

Zainul A, Noehi N (1997). Assessment of Learning Outcomes. Jakarta: PAU-PPAI.