The characteristics of final mathematics test items based on classical test theory

The characteristics of final mathematics test items based on classical test theory

A T Panjaitan[1*] and H Retnawati[2]
[1]Department of Mathematics Education, Graduate School Program, Yogyakarta State University, Indonesia
[2]Department of Mathematics Education, Yogyakarta State University, Indonesia

*Corresponding author: agnesteresa.2018@student.uny.ac.id

Abstract. Final Examination is the part of an evaluation that exists in school to assess student's knowledge. This research aimed to show the characteristic of the final mathematics test in Junior High school, grade 8. This research was an explorative descriptive using a quantitative approach. Data collection was gathered using the documentation of mathematics' test in Binjai, North Sumatera including student's answers. The characteristics of test items were analyzed by expert judgment using the classical test theory approach. The average of content analysis was 0.81, there were three items without key answers so it can not be analyzed. The reliability average was 0.84 which has very good quality and difficulty index indicated the difficulty level of questions was on three categories (easy, moderate, and difficult). The difficulty is in the cartesian position, the relation of two sets, determining gradient, determining the solution of linear equation systems of two variables.

1. Introduction

Assessment in education has been an interesting topic at many levels. There are many kinds of assessments that the teacher used to evaluate a student's mathematical learning. Not only for evaluating student's mathematical learning but also for providing student grades, national accountability, tracking on the right system, allocating resources within a district, regulating involvement, upgrading the class' activities, or giving suggestion and response to students and their guardians [1]. Assessing performance is the influential activity for any subjects, it will facilitate formative report to the pupil and educator to guide future study and occasional summative feedback [2]. Therefore, to present the real achievement of student's learning, it is essential to give the best qualification of the test [3].

NCTM [4] noted that an assessment instrument containing many computational items and relatively few problem-solving questions is poorly aligned with a curriculum that stresses problem solving and reasoning. It also highlights the integration of mathematical knowledge must contain tasks, mathematical power, a task with a non-unique solution. Setting the standard of the tests is not as simple as the expectation, there are some factors to make the good qualification of tests such as the technique chosen by test makers.

Moreover, the criterion shown by the score from the other examination results is similar. It can be concluded that the test can be replaced by other tests, such as school performance tables also impose the same requirement. Besides, the main purpose of the public examination is to provide information for future educational system and vocational selection decisions, and it is a feature of this selection that information from different examinations is combined and compared. As a result, it is critical to

arrange the good qualification of questions, with regard to the judicial impartiality which required examination as the selection, to show the same common scale at the similar point on other kind of tests in the same level [5]. Therefore, it is needed to put on the valid score for any examination the students face.

Thorndike [6] stated that the implementation of measurement error diagnosis by discrepancy measurement. the reliabilities of the tests, their correlation, and the size of the score difference involved as the aspects of the tests. Furthermore, Thorndike mentioned that it is clearly impossible to develop the accurate diagnostic interpretation of score differences by the existing instrument. If the diagnostic interpretations used correctly by the educators, it would affect the learning process. The result of the score reports the learning process and the way of student's thinking. Thus, the test must have been considered as valid and in the good categories. The importance of the valid test is not only to monitor the students learning but also to enhance the learning method, the proposition of items, and the standard of any selection.

One of the best ways to know the quality of the tests is by using the analysis of item characteristics [7]. In general, the analysis of item characteristic facilitates the teacher to know how well-qualified the test is [8]. Lamentably, the researchers have found that one of the junior high schools in Binjai designed the test without analyzing the characteristic of the items for Mathematics Final Examination. The tests were designed by multiple choice tests. The quality of the good test such as the discrimination index. Index of difficulty, distractor effectiveness has not known. However, the characteristic of a good test should be developed by the teacher.

The test item discrimination index shows the ability of an item in recognizing the participants having the high grades or the low grades, the test item discrimination index distinguishes the students who do the test well or not. Hence the discrimination index is divided into three categories: positive, negative, and zero. The difficulty index describes the percentage of students answered correctly [8]. The quality of the test can be found after the analysis has been done. Based on the literature review, the analysis of item characteristics will show the qualification of the test. Accordingly, this research aims to show the quality of The Junior High School Mathematics school examination test in Binjai. Furthermore, this research is expected to develop a well-qualified test instrument of mathematics in the future.

2. Method

The research was a document analysis of the descriptive quantitative approach. Due to the existence of the data before the research, the data functioned as the secondary data. The data used for this study was the mathematics final examination with the answer sheets. The multiple choice test had been answered by 100 students in the 8th grade. There are four choices, one of them will be the key answer while the others became the distractors. The procedures might be described as the following:

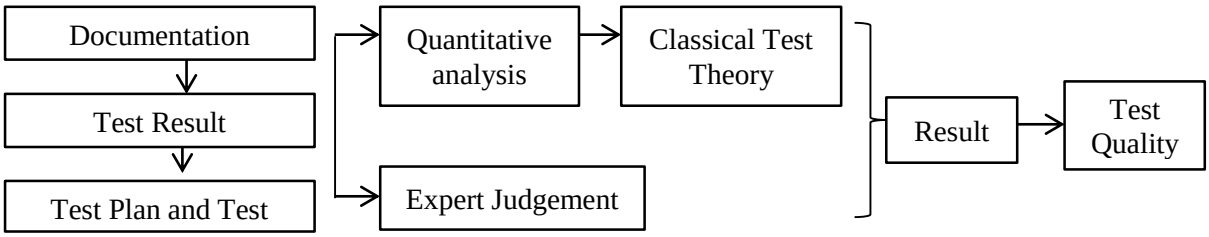

Figure 1. Research procedures

After collecting the documentation, The first analysis regarding to the test material will show the validity of the content. Allen [9] stated content validity is denoted through a rational analysis of the content of a test and based on individual, subjective judgement. According to Aiken [10] to determine the validity index as follows:

$$V = \frac{\sum s}{n(c-1)}$$

The index of validity of the test item is termed by V. the experts' assigned score minus the lowest score in that category is symbolized as S (where: s= r-r0, r is the assigned score by experts and r0¬ is the lowest score in the category). n is the number of experts and c is the number of categories that might be selected by an expert. The quantitative analysis approach to classical test theory consisted of reliability, discrimination index, and distractor effectiveness. According to Miller, Linn & Gronlund [11], The reliability index category should be based on the correlation coefficients in Table 1.

Table 1. Reliability criteria

Reliability Index	Criteria
0.81 - 1.00	Very Good
0.61- 0.80	Good
0.41- 0.60	Quite
0.21- 0.40	Poor
0.00 – 0.20	Very Poor

The discrimination index that the researcher implemented was the biserial correlation point. Mardapi notes the discrimination index can be presented in the following table [12].

Table 2. Discrimination index criteria

Discrimination index	Criteria
> 0.30	Good and Acceptable
0.20 – 0.30	Quite Good and need repairing
< 0.20	Not Good and not acceptable

And, the measurement of difficulty index test items can be seen in the table below.

Table 3. Difficulty index criteria

Difficult index	Criteria
> 0.70	Easy and Not Good
0.30 – 0.70	Medium and Good
< 0.30	Hard and not Good

According to Attali & Bar-Hillel [8] the distractor is known as qualified, were it selected at least by 5 % of all the participants' test and had a negative in biserial correlation point selection. Thus, if the distractors are said as inaffective, it should be changed by the others, then the mastered students would be attracted to select the options.

3. Result and discussion

3.1. The qualitative analysis based on the quality of the test

The results of the test plan qualitative analysis for every question of Junior high School Mathematics Examination in Binjai had grouped as Quite Good and Good. The results provided by the table below:

Table 4. Qualitative analysis result

Criteria	Number of items
Good	1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14 ,15, 16, 17, 19, 20, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34
Quite Good	8, 9, 18, 21, 26, 35

The result above showed that there had been 82.85 % items of the 35 numbers have good quality and 17.15 % items test that had quite good quality with minor revisions in terms of indicator suitability).

3.2. The quantitative analysis based on the quality of the test

The validity test was conducted by correlating the test to the other standardized test using the QUEST program. Concerning to the criteria validity test, the Aiken index proved the mean of reliability is equal to 0.84 which means The Very High category, bigger than 0.81 (>0.81). Based on the output generated by the QUEST program, the discrimination index for every item of Junior High School Mathematics Examination conducted in Binjai can be read from the correlation point biserial. The discrimination index would be presented in the table below:

Table 5. The discrimination index

Criteria	Number of Items
Good	7, 9 , 10, 11, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29 , 32, 34
Quite Good	3
Not Good	1, 2, 4, 5, 6, 8, 12, 14, 24, 30, 31, 33, 35

The discrimination index average score of 35 items belongs to the "Good category" (0,3). The difficulty index of every Junior High School Mathematics Examination might be seen from the proportion of correct answer in the following table:

Table 6. The difficulty index

Criteria	Number of Items
Easy	1, 2, 5, 12
Medium	3, 4, 6, 7, 9, 10, 11, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 27, 29, 32, 35
Hard	8, 17, 24, 25, 26, 34

Based on The difficulty index of every Junior High School mathematics Examination item, four items belong to the easy category, it is around 11. 42% among all. The indicator for the question is determining the number pattern, fibanocci sequence, and locating the cartesian coordinate. The medium category belonged to 21 items among 35, it was about 60% item is Medium. It was represented by five topics examined such as number pattern, cartesian diagram, relation and function, equation of the straight line, linear equation in two variables. The five materials used in the test were in the medium category at most. The hard category for the items belonged to 6 numbers. The questions determined about relation and function, equation of the straight line, linear equation in two variables. It means 17.14% of the items are hard for the students.

Instead of all the categories, there are 3 questions without any key answer. The numbers are 30, 31, 33. The questions can be seen below:

30. The price of one dress and three clothes is IDR.185.000,- and the price for three dresses and three clothes for the same label is IDR.275.000. Then, the price of 2 dresses and one cloth is
a. IDR. 107.000,- c. IDR. 170.000,-
b. IDR. 115. 000,- d. IDR. 195.000,-

The indicator for this question is solving the problems about the Linear Equation in Two variables in real life. The price of a dress is IDR.46.666 and the cloth is IDR. 46.000. Then, the price of 2 dresses and one cloth is IDR. 136.666.

The other question is in the same topic by another kind of the linear equation in two variable application.

31. If the sum of the two numbers is 72 and the difference is 6. The multiplication of those numbers is
a. 429 b. 249 c. 78 d. 43

The reason why the solution cannot be found is the answer must be 1287, due to the value of x is 39 and the value of y is 33. The last number is still in the same topic, yet the question is about finding the value of the linear equation in two variables, after finding the value of x and y from the question. The reason why the solution cannot be found is the answer must be 1287, due to the value of x is 39 and the value of y is 33.

The last number is still in the same topic, yet the question is about finding the value of the linear equation in two variables, after finding the value of x and y from the question. The other question is in the same topic by another kind of the linear equation in two variable applications.

31. If 3x-y = 15 dan x + y = 3. Thus the result of 2x – y is equal to
a. 10 b. 12 c. -10 d. 15

The solution of the question above is -19,5 where the x = 4.5 and y = 28,5. To find the result of the last question participant should substitute the value of each variable. Then, after finding the result there is no answer to be found.

3.3. Discussion

Based on the analysis of classic theory, The Junior High School Mathematics Test in one of school in Binjai has 60% medium category among 35 questions. This result is one of the purposes of the analysis item characteristic. It can present the quality of the questions that can enlarge the quality of the test by revising the "not good" category to be a good category. Then the questions can be in the good category for all items [8]. Three of the questions cannot be solved because the choice did not provide the answer. This result cannot be said as an adequate qualification of the test. Moreover, the examination is viewed as the report of students mathematical thinking.

The Difficulty index for the test has the representative for five subjects that were used. The distractor effectiveness from the table above shows for the easy category, each distractor was not chosen by most of the participants. For the medium category, The distractor effectiveness shows each distractor has more probability chosen by the participants. For the hard category, the distractor has the most number chosen by the participants compared to the key answer.

Kriswantoro [13] states that to make meaningful questions, the teacher should consider the distractors for all items. The items with "not good" distractors have the biggest chance to answer by guessing due to the four choices the participants have. Conversely, the distractor that is close enough to the answer will make the question in the hard category. Furthermore, the existence of the unlogic distractor can make the students decide the correct answer.

According to The discrimination index of the test can conclude as a good category. Due to 19 questions belonged to the good category, one is good enough and 13 is the "not good" category. The three questions without answer key cannot be measured. Some factors that put the "not good" category are the difficulty index is too easy and too difficult for students [14].

The analysis of classic theory can increase the quality of the test by checking the indicator of the learning with the items so the question is able to evaluate the student's learning process. The questions without the answer will increase the probability for the students for guessing the answer. Therefore, the teacher should check the distractor and the questions for the test. Moreover, the distractor must be applicable considering Mathematics can be implemented in daily activities[15]. Nevertheless, there are the other factors that affect the examinee's measurement such as the internal(pyshcology) and external (environment) factors [16].

4. Conclusions

The result shows the content analysis(aiken index) is 0.81 and the reliability of the test is 0,84. There are three questions without the key answer. Thus, the questions that can be analyzed in this research are 33 questions. Six of the question considered a hard category that represents four topics of five topics that are used in the test. The distractor work mostly for the hard category. Based on the quantitive analysis by employing classical test theory approach, all of the test items in the Junior High School Mathematics Examination in Binjai had a good category. Nonetheless, the three questions

without the answer key should be revised by the teachers to make the test is adequate to be answered by the participants.

According to the studies described previously, theoretically, multiple intelligences approach in Problem-Based Learning has the potential to improve students' mathematical literacy skills. By implementing Problem-Based Learning methods, students can actively develop their skills and abilities in solving mathematical problems through the stages of Problem-Based Learning. In addition, students are able to develop their own mathematical concepts through group discussions. Teaching by paying attention to each student's intelligence and characteristics can also help teachers manage the class well. Learning can be tailored to the students' interests and needs as multiple intelligences approach can enhance each type of students' intelligence.

References

[1] Newton P 2007 *Assessment in Education Journal.* (Informa UK limited) pp 149-170
[2] Swan M and Burkhardt H 2012 *Journal of the International Society of Design and Development in Education* (*Educational Designer volume 2*) pp 1-41
[3] Kartowigran B, Munadi S, Retnawati H, and Apino E 2017 *SHS Web of Conferences (GC TALE vol 42)* p 6
[4] NCTM 1989 *Curriculum and Evaluation Standards for School Mathematics.*
[5] Creswell J W 2015 *Penelitian kualitatif dan Desain Riset.* (Yogyakarta: Pustaka Pelajar).
[6] Bejar I 1984 *Journal of Educational Measurement.* **21** pp 175-189
[7] Santoso A 2018 *Pendidikan Matematika dan Sains Journal* (Yogyakarta State University) **6** p 159
[8] Sutrisno 2016 *Riset Pendidikan Matematika Journal* (Yogyakarta State University) pp 162-177
[9] Allen M J and Yen W M 1979 *Introduction Of Measurement theory* (California USA: Brooks/Cole Publishing Company).
[10] Aiken L R 1980 *SAGE Journal* **40** pp 955-959
[11] Miller M D, Linn R L and Gronlund N D 2009 *Measurement and Assessment in Teaching* (New Jersey USA: Pearson).
[12] Mardapi D 2012 *Pengukuran penilaian & evaluasi pendidikan* (Yogyakarta: Nuha Litera).
[13] Amelia R N and Kiswantoro K 2017 JKPK journal **2** pp 1-12
[14] Thorndike 2005 *Measurement and Evaluation Education* 2005 (New Jersey USA: Pearson Education).
[15] Maulina Y and Retnawati H 2017 *ICRIEMS Proceedings of fifth annual conference* (Yogyakarta State University) p 467
[16] Bichi, Ado A 2016 *IJSS journal* **2** p 32

Acknowledgements

Special thanks should be given to Jesus Christ, without His blessing, the process of this research would not have been completed and special thanks should and always be given from the author to the sponsor of the author's education funding, LPDP. This research would also never have been possible without the support and guidance of various people at the Yogyakarta State University. Mrs. Dr. Heri Retnawati as a lecturer who gave ideas, advice and moral support in writing this article. Members of my class, thanks for being one of my support systems and giving extra information about Quest Program. Further, the author would like to express my appreciation to Ms. Naila Muroda, S.Pd who became my proofreader for checking the English grammar in this writing. Lastly, a special note of thanks should also be given to my parents and my sisters, Sonia Natasya Panjaitan and Sintya Monica Panjaitan. Your unconditional love, devotion, and optimism was greatly appreciated.

.