JURNAL RISET PENDIDIKAN MATEMATIKA
Volume 3 - Number 2, November 2016, (162 - 177)
Available online at http://journal.uny.ac.id/index.php/jrpm

AN QUALITY ANALYSIS OF THE MATHEMATICS SCHOOL EXAMINATION TEST

Hadi Sutrisno
SMP Negeri 1 Tanahmerah Bangkalan. Jalan Raya Tanah Merah No.105, Tanah Merah Dajah, Bangkalan, Kabupaten Bangkalan, Jawa Timur 69172, Indonesia
Korespondensi Penulis. Email: math.united@gmail.com

Received:10th December 2016; 10th December 2016; Accepted: 10th December 2016

Abstract

The study is to describe: (1) the quality of the Junior High School Mathematics School Examination Test for the 2015/2016 Academic Year in Kabupaten Bangkalan based on the test item qualitative analysis, (2) the quality of the Junior High School Mathematics School Examination Test for the 2015/2016 Academic Year in Kabupaten Bangkalan based on the test item quantitative analysis, and (3) the Junior High School Mathematics School Test equating for the 2015/2016 Academic Year in Kabupaten Bangkalan. A test is said to be qualified if the test fulfills the criteria of validity, reliability, and good characteristic. A test is said to be equivalent if the scores of the test that has been conducted might be exchanged to those of the other test. The data were taken from the school examination script complete with the students' answer sheets. The qualitative data analysis was conducted by means of the expert judgement. On the other hand, the quantitative data analysis was conducted by means of the Classical Test Theory by Iteman and the Item Response Theory by BilogMG. These programs were implemented in order to define the test quality quantitatively. Then, in order to analyze the equivalence among the test the series, the researcher implemented item-characteristic curves. These curves were drawn by means of Geogebra. The results of the study have shown that: (1) qualitatively, the quality of mathematics school examination test plan is quite good while the school examination quality is quite good but not so good; (2) quantitatively, the school examination test quality is good, and (3) for the test equating, based on the item-characteristic curves the school examination tests are equal.

Keywords: test quality, qualitative analysis, quantitative analysis, tests equating, classical test theory, item response theory, test characteristic curve.

How to Cite: Sutrisno, H. (2016). An analysis of the mathematics school examination test quality. *Jurnal Riset Pendidikan Matematika, 3*(2), 162-177. doi:http://dx.doi.org/10.21831/jrpm.v3i2.11984

***Permalink/DOI*:** http://dx.doi.org/10.21831/jrpm.v3i2.11984

INTRODUCTION

Assessment is an important part in ensuring quality of educational outcome. According to Budiman & Jailani (2014, p.140), the quality of learning outcome assessment instruments directly affects the accuracy on the status of students' learning outcomes achievement. Meanwhile, according to the NCTM (2000, p.22), good assessment might increase the learning process of learners. One of the characteristics for a good assessment is that the good assessment makes use of a good measuring tool, which is able to convey the message to learners. The measuring tool which is commonly used in educational assessment is the test. Regarding the test, Phopam (2009, p.42) states that test is an important measurement tool for assessing the quality of learning. The sensitivity of a test that has been designed by the teachers is implemented in order to look for the evidence on the effectiveness of the learning process that has been carried out.

A well-qualified test is designed based on the test-designing procedures. Allen & Yen (1979, p.118) states that all who designs tests should pay attention to the organization test procedures. Unfortunately, the Subject Teachers Forum (MGMP, *Musyawarah Guru Mata Pelajaran*) for the Junior High School in Bangkalan as the compiling team of Junior High School Mathematic Examination Test for the 2015/2016 Academic Year has given less

Print ISSN: 2356-2684, Online ISSN: 2477-1503

attention to the test procedure organization. Based on the interview with the Head of Junior High School/Senior High School Curriculum in Bangkalan on August 20th, 2015 the researcher has found that from one year to another the compilers of Junior High School Mathematic Examination Test only designed the tests without any item analysis.

The test that has been administered into the the Junior High School Mathematic Examination Test in Bangkalan is the multiple choice one. The multiple choice test-design has been selected because the test participants are more motivated in completing the multiple choice test items rather than the essay test items. In relation to the problem, Jailani & Retnawati (2016, p.5) states that the test participants tend to be lazy, be unconfident, and be difficulty to complete the essay test item. Similarly, according Nitko & Brookhart (2011, p.166), a multiple-choice item consists of one or more introductory sentences followed by a list of two or more suggested responses which is easier to comprehend. In completing the multiple choice test design, the test participants must choose the correct answer among the options of the answers that has been provided.

The Junior High School Mathematic Examination Test in Bangkalan has been designed without analyzing the produced test which quality and equality has not known. According to Popham (2009, p.51), the usefulness of an educational test for particular assessment functions should be judged according to the following four factors: reliability, validity, bias and instructional sensitivity. Similarly, Salvia, Ysseldyke, and Bolt (2010, p.141) state that teachers should develop technically adequate assessment procedures. Two aspects of this adequacy are especially important: content validity and reliability. Therefore, the teachers should develop a test based on the assessment procedure. The two important aspects of the tests that have been developed are validity and reliability. With regards to the validity and reliability, Nurlita (2015, p.42) states that the characteristics (validity, reliability, discrimination index and difficulty index) of a good test should be achieved. In other words, a test is said to be well-qualified if the test has good validity, reliability and other characteristics. The other characteristics of a good test, then, might include discrimination index, index of difficulty, distractor effectiveness, model of fit and guessing.

In order to define the quality of Junior High School Mathematic Examination test in Bangkalan, an analysis should be conducted in order to make sure that the examination test that has been implemented is able to provide information about the quality of each item on the test. The analysis should be conducted by means of qualitative and quantitative analysis. The test item qualitative analysis should be conducted by the experts. On the contrary, in conducting the tes item qualitative analysis toward the multiple choice test design the teachers might implement the following criteria:

Table 1. Qualitative Analysis Criteria of Multiple Choice Test Item

Aspect	Criteria of Analysis
Material	Matching the item and learning indicator
	Matching the item and learning targets
	The correct answer to one item is independent of the correct answers the other item
	Each alternative answer is plausible
	There is only one correct answer
Construction	Pictures, graphs, tables and sentences are understandable
	Avoid using negative words in the stem
	Make the stem as brief as possible
	Place alternative answer in logical or numerical order
	All of the alternatives answer are homogeneous
	Avoid using "all of the above" or "none of the above" as much as possible
Language	Matching the item and the Bahasa Indonesia rules
	The vocabulary and sentence structure are at a relatively understandable

According to Miller, Linn & Gronlund (2009, p.150), there are some checklists that might used in reviewing the test gratings. The checklists are as follows: (1) whether the tests plan are suitable with purpose of the test; (2) whether the tests plan show domain competence to be measured; (3) whether the gratings test shows the learning outcomes which will be measured; (4) whether the test item plan measures more than one purpose of learning; (5) whether the test format in the grating test is suitable to the learning outcomes that will be measured; (6) whether the indicators in the tests plan might be made on the item tests; (7)

Print ISSN: 2356-2684, Online ISSN: 2477-1503

whether the test item in the tests plan has represented the desirable competence; and (8) whether all the tests plan are suitable with the results that will be desired. Based on the above opinion the researcher would like to concude the criteria for the grating test assessment. The criteria for the grating test plan are: (1) the grating test plan should correspond to the learning objectives; (2) the grating test plan should display the competence that will be achieved; and (3) the grating test plan should be easy to understand.

A qualitative analysis toward the test material will produce the content validity. According to Allen & Yen (1979, p.95), content validity is established through a rational analysis of the content of a test and its determination is based on individual, subjective judgment. In order to assess the expert agreeement in proving the content validity, the researcher will implement the validity index is. According to Aiken (1980, p.956) and Retnawati (2016, p.18) the formula for determining the validity index is as follows:

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

V is the index validity of the test item. s is the assigned score by experts minus the lowest score in that category ($s = r - r_0$, where r is the assigned score by experts and r_0 is the lowest score in the category). n is the number of experts. Last but not the least c is the number of categories that might be selected by an expert.

In addition to the content validity, another validity that will be required within a test is the criteria validity. For determining the validity criteria in the study, the researcher will implement the predictive validity. According to Ary, Jacobs, Sorensen, & Razavieh (2010, p.229), predictive validity evidence is the relationship between scores on a measure and criterion scores available at a future time. For obtaining predictive validity coefficients the researcher will implement the coefficient of correlation between the test scores and the score criteria; the coefficient of correlation will be the clues for the relationship between test scores with the score criteria. According Urbina (2014, p.207), in order to obtain the validity coefficient (r_{xy}) a researcher should implement the product moment correlation as follows:

$$r_{xy} = \frac{\sum xy}{(N-1)(SD_x)(SD_y)}$$

r_{xy} is correlation of product moment. N is the number of test participants. SD_x is the standard deviation of test scores. Finally, SD_y is the standard deviation of score criteria.

There are two approaches to the quantitative analysis test, namely the classical test theory and the item response theory. According Mardapi (2012, p.198), the classical test theory is a theory that makes use of the simple mathematic model in order to show the relationship among the observation score, the actual score and the error score. The assumptions in classical test theory might be developed into various formulas that are useful for making measurements. The resulting formulas from the classical test theory will be the characteristics of test items such as reliability, discrimination index, the index of difficulties and distractor effectiveness. Meanwhile, according to DeMars (2010: p.3) the item response theory (IRT) models is shown by the relationship between the ability or the trait (symbolized by θ) that has been measured by the instrument and the item response. Similarly, Retnawati (2014, 93) states that equalization is a process of linking the test scores that have been statistically and conceptually intended to be interchangeable. In short, the item response theory is a model that shows the relationship between the ability or the trait (symbolized by θ) as having been measured by the instrument and the response item.

The quantitative analysis approach to classical test theory test that serves to determine the test characteristics includes reliability, discrimination index, index of difficulty and distractor effectiveness. On the other hand, the quantitative analysis approach to the item response theory approach that serves to investigate the test includes model of fit discrimination index, index of difficulty and guessing.

Reliability is a coefficient of correlation that shows the test power in terms of consistency within the measurement test results. A test is said to have a high reliability if the test provides the consistent results. According Faremi (2016, p.60), reliability is all about the consistency, stability, dependability and predictability of any research instrument or test which can be estimated using test-retest, split-half, parallel/equivalent, KR-2021 and cronbach alpha.

A test item discrimination index refers to the ability of an item in distinguishing the test takers who have high grades and the test takers

Print ISSN: 2356-2684, Online ISSN: 2477-1503

who have low grades. The test item discrimination index is based on the opinion of Kubiszyn & Borich (2003, p.198): discrimination index measures the extent to which a test item discriminates or differentiates between students who do well on the overall test and those who do not do well on the overall test. Discrimination index is a test item characteristics that distinguish between test takers who answered all the test well and which are not. Discrimination index is divided into three categories: positive, negative and zero.

Difficulty index, on the other hand, refers to the proportion of test takers who respond to the test items correctly. The difficulty index is based on the opinion by Chauhan (2015, p.1608): "difficulty index also called ease index, describes the percentage of students who correctly answered the item". The difficulty index also describes the percentage of students who answered test items correctly.

After the test characteristics have been found, the test quality will also be found. Then, the equality test will be known as well. According to von Davier (2011, p.23), equating is the strongest form of linking between the scores on two tests. Equating may be viewed as a form of scale aligning very strong in which requirements are placed on the tests being linked. Equating is the best form that links the scores to the two tests. Equating might also be seen as a form of test scale allignment that has very good relationship within the similar tests.

Based on the explanation, the study is to describe: (1) the quality of the Junior High School Mathematics School Examination Test in Bangkalan; and (2) the Junior High School Mathematics School Test equating in Kabupaten Bangkalan.

METHOD

The study was a document analysis that made use of descriptive quantitative approach. The study was conducted on April-May 2016 (the implementation of Junior High School Mathematics Examination for the 2015/2016 Academic Year in Bangkalan). The research site then was on the junior high schools around the Regency of Bangkalan area.

The population in the study was the participants of Junior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan. The participant number was 6,575 respondents who had been selected from 56 junior high schools. The sampling technique that had been used was the proportional stratified random sampling. The sample number in each stratum was determined based on the Kricjie Table. The results of sample selection were presented in Table 2 as follows.

Table 2. The Number of Samples

Strata	School Number	Subject Number	Sample Number
City	22	3601	656
Non City	34	2974	477

The setting that had been implemented in the study was the field study. Field study is a research setting that tests a variety of factors within natural conditions in which the activity takes place normally and almost no involvement of researcher. In order to maximize the setting, the researcher designed several procedures. The procedures might be described as following:

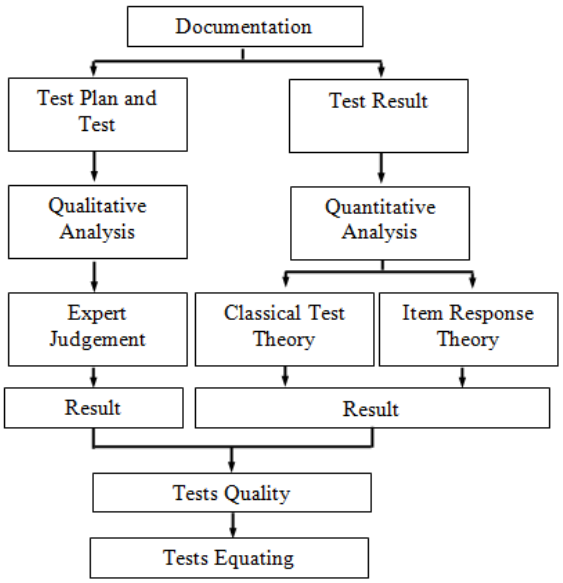

Figure 1. Research Procedures

The necessary data for the study were collected by means of documentation. The documentation technique had been implemented in the study in order to collect the documents related to the instrument and the answer sheet of Junior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan.

The data analysis techniques that would be implemented in the study included the qualitative analysis, the quantitative analysis and the equating analysis tests. The qualitative analysis included the gratings qualitative analysis and the Junior High School Mathematics Examination in the Regency of Bangkalan. The criteria of test

Print ISSN: 2356-2684, Online ISSN: 2477-1503

quality based on the qualitative study might be shown in Table 3.

Table 3. Test Quality Criteria Based on Qualitative Analysis

Criteria	Item Test Aspect	Test Plan Aspect
Good	The test items have all material, construction and language aspects	The test items have the following aspects: capability to match the test plan to the learning targets, the capability to cover the important competences and the capability to make the test plan understandable
Quite Good	The test items have all material aspects	The test items have the following aspect: the capability to the test plan to the learning targets
Not Good	The test items do not have one of the material aspects	The test items do not have the following aspect: the capability to match the test plan to the learning targets

Then, the quantitative analysis included the evidence of validity, the approach of classical test theory and the item response theory. The evidence of validity included the content validity and criteria validity. After having attained the content validity and the criteria validity, the researcher would like to conduct the categorization. According to Urbina (2014, p.208), the validity category would be provided as in the following Table 4.

Table 4. Validity Criteria

Validity Coefficients	Validity Criteria
0.40 – 1.00	Acceptable
0.00 – 0.39	Not Acceptable

Next, the analysis by means of classical test theory approach included the analysis of test characteristics that consisted of: reliability, discrimination index, difficulty index and distractor effectiveness. For the test reliability estimation, the researcher made use of the Kuder-Richardson Formula 20 (KR-20). According to Miller, Linn & Gronlund (2009, p.110), the reliability index category should be based on the correlation coefficients in Table 5.

Table 5. 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 method that the researcher implemented in estimating the discrimination index was the biserial correlation point. According Mardapi (2005, p.5) and Ebel and Frisbie (1991, p.232), the determination toward the functioning of a discrimination index test items should be as provided in the following Table 6.

Table 6. Discrimination Index Criteria

Discriminat Index	Criteria
> 0.30	Good and acceptable
0.20 – 0.30	Quite good and need repairing
< 0.20	Noot good and not acceptable

The method that the researcher implemented in estimating the difficulty index was the proportion of correct answer. According to Allen & Yen (1979, p.121) and Mardapi (2012, p.186), the determination of difficulty index test items should be as provided in the following Table 7.

Table 7. Difficulty Index Criteria

Difficulty 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 (2003, p.123), distractor is said to be effective if it was chosen at least by 5% of all the participants test and have a negative biserial correlation point. Ineffective distractor should be replaced with others that may be more interesting for participants who have not mastered the knowledge in test items to choose the distractor.

The second quantitative analysis that the researcher implemented was the item response theory approach. The criteria for the test item based on the item response theory by Hambleton & Swaminathan (1985, p.36), and Gunartha, Kartowagiran, & Suardiman (2014, p.36) were as follows:

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Table 8. Item Response Theory Criteria

PL	Criteria		
	Good	**Quite Good**	**Not Decision**
1-P**L**	$p > 0.05$ $-2.00 \leq b \leq 2.00$	$p > 0.05$ $b < -2.00$ or $b > 2.00$	$p < 0.05$
2-P**L**	$p > 0.05$ $0 \leq a \leq 2.00$ $-2 \leq b \leq 2.00$	$p > 0.05$ Have not *a* or *b* criteria	$p < 0.05$
3-P**L**	$p > 0.05$ $0 \leq a \leq 2.00$ $-2.00 \leq b \leq 2.00$ $c \leq 0.25$	$p > 0.05$ Have not *a, b* or *c* criteria	$p < 0.05$

To determine the test equating the research wold implement the test characteristic curve. According to Retnawati (2015, p.279), the two packages of tests would be equal if the characteristic curves from two test packages had been adjacent.

RESULTS AND DISCUSSION

The Quality of the Tests Based on the Qualitative Analysis

The results of the test plan qualitative analysis for each package examination test Junior High School Mathematics Examination in Bangkalan were as follows:

Table 9 showed that test plan of Junior High School Mathematics Examination in the Regency of Bangkalan had fallen into the "Quite Good" or "Fit for Use" with revision.

The results of qualitative analysis toward the material aspects, the construction and the language for each package Junior High School Mathematics Examination that consisted of 40 items test showed that there had been some tests which did not possessed each criteria. The complete results from the analysis toward qualitative aspects of material, construction and languages were presented in the following Table 10.

Table 9. Test Plan Qualitative Analysis Result

Package	Result		
	Rater 1	**Rater 2**	**Rater 3**
41	Quite	Quite	Quite
42	Quite	Quite	Quite
43	Not good	Quite	Quite
44	Good	Quite	Quite
45	Good	Quite	Quite

Table 10. 41 Package Qualitative Analysis Result

Criteria	Number of Items
Good	-
Quite Good	2, 3, 5, 6, 11, 13, 18, 22, 23, 24, 25, 27, 28, 30, 31, 33, 34, 35, 36, 38, 39, 40
Not Good	1, 4, 7, 8, 9, 10, 12, 14, 15, 16, 17, 19, 20, 21, 26, 29, 32, 37

Table 11. 42 Package Qualitative Analysis Result

Criteria	Number of Items
Good	18, 25, 33, 34, 36
Quite Good	2, 3, 5, 6, 11, 12, 13, 14, 21, 22, 23, 24, 27, 28, 31, 35, 38, 39, 40
Not Good	1, 4, 7, 8, 9, 10, 15, 16, 17, 19, 20, 26, 29, 30, 32, 37

Table 12. 43 Package Qualitative Analysis Result

Criteria	Number of Items
Good	10
Quite Good	1, 2, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 24, 25, 26, 27, 28, 29, 30, 31, 36, 38, 40
Not Good	3, 4, 22, 23, 31, 32, 33, 34, 35, 37, 39

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Table 13. 44 Package Qualitative Analysis Result

Criteria	Number of Items
Good	17
Quite Good	2, 6, 11, 14, 15, 16, 18, 19, 22, 24, 25, 29, 30, 32, 35, 37, 39
Not Good	1, 3, 4, 5, 7, 8, 9, 10, 12, 13, 20, 21, 23, 26, 27, 28, 31, 33, 34, 36, 38, 40

Table 14. 45 Package Qualitative Analysis Result

Criteria	Number of Items
Good	-
Quite Good	2, 6, 9, 11, 13, 15, 16, 17, 18, 19, 21, 25, 28, 31, 32
Not Good	1, 3, 5, 7, 8, 10, 12, 14, 20, 22, 23, 24, 26, 27, 29, 30, 33, 34, 35, 36, 37, 38, 39, 40

The results displayed in the Table 10 indicated that there had been 45.00% items of the 41 packages that did not have good quality in terms of material. Then, there had been more than 55% items test that had quite good quality with minor revisions in terms of construction and language.

Next, the results displayed in the Table 11 showed that there had been 12.50% items of the 42 packages that did not have good quality in terms of material, construction and language. On the other hand, there had been 47.5% items that had quite good quality with minor revisions in terms of construction and language. The remaining 40% of these items did not have good quality because these items did not fulfill the material criteria (both in terms of indicator suitability and of the use of alternatives).

Furthermore, the results displayed in the Table 12 showed that there had been 2.50% items of the 43 packages that did not have good quality in terms of of material, construction and language. On the contrary, there had been 70% items that had quite good quality with minor revisions in terms of construction and language. The remaining 27.50% the items did not have good quality because these items did not fulfill the material criteria (in terms of indicators suitability, of incompatibility with the test objective and the use of alternatives).

The results displayed in the Table 13 showed that there had been 2.50 % items of 44 packages that had good quality. Then, there had been 42.50 % of these items that had quite good quality with minor revisions in terms of construction and language. The remaining 55.00% of these items did not have good quality because they did not fulfill the material criteria (both in terms of indicators suitability or and of the use of alternatives).

Last but not the least, the results displayed in the Table 14 showed that there had been 37.50 % of these items had good quality with minor revisions in terms of construction and language. The remaining 62.50 % of these items did not have good quality because they did not fulfill the material criteria (both in terms of either indicators suitability and of the use of alternatives).

The Test Quality Based on the Quantitative Analysis

Evidence-Based on Content Validity

The evidence-based on content validity proved the test validity toward the test material by the experts. In the study, the researcher and two experts performed the evidence-based content validity test within the mathematic evaluation. The results of the evidence-basec content validity test toward the Junior High School Mathematic Examination in the Regency of Bangkalan for each package would be provided as follows.

Table 15. Contents Validity Result

Pack	Averg Rater 1	Averg Rater 2	Averg Rater 3	Σs	V
41	4.30	4.80	4.75	10.85	0.90
42	4.55	4.93	4.75	11.23	0.94
43	4.25	5.00	4.93	11.18	0.93
44	3.73	4.40	4.58	9.70	0.81
45	3.45	4.98	4.55	9.98	0.83

Based on the results displayed in the Table 15, the researcher found that the content validity for each package for the Regency of Bangkalan Area belonged to the "Acceptable" category. The content validity might be seen from the test index validity (V) for each package that had been bigger than 0.40 ($V > 0.40$).

Evidence Based on Criteria Validity

The evidence-based on criteria validity test was conducted by correlating the test to the other standardized tests. In relation to the evidence-based on criteria validity test, the study

Print ISSN: 2356-2684, Online ISSN: 2477-1503

applied the predictive validity proof in which the Junior High School Mathematic Examination served as the predictor and the Junior High School Mathematics National Examination served as the criteria. The evidence-based on criteria validity test made use of the correlation coefficient (validation coefficient) between the scores of the school test and those of national the examination. The results of the predictive validity evidence would be displayed in the following table.

Table 16. Predictive Validity Result

Package	Validity Coefficient (r_{xy})
41	0.525
42	0.528
43	0.555
44	0.485
45	0.496

Based on the results displayed in the Table 16, the researcher found that all packages of the Junior High School Mathematics Examination for the Regency of Bangkalan area had been accepted in terms of proving the criteria validity. The reason was that the validity coefficient of each package in the validation test had been bigger than 0.40 (> 0.40).

The Test Quality Based on the Classical Test Theory

Based on the estimates generated by the KR-20 techniques for the analysis of each test package in the Junior High School Mathematics Examination, the researcher obtained the reliability index as follows.

Table 17. Reliability Index

Package	Reliability Index
41	0.917
42	0.921
43	0.907
44	0.913
45	0.910

Based on the results displayed in the Table 17, the researcher found that all packages in the Junior High School Mathematics Examination within the Regency of Bangkalan area belonged to the "Very High" category in terms of reliability estimation. The reason was that value of reliability index of each test package had been bigger than 0.81 (> 0.81).

Based on the output generated by the Microcat Iteman software, the discrimination index for each pakacged of Junior High School Mathematics Examination conducted in the Regency of Bangkalan might be seen from the correlation point biserial. The discrimination index in each test package would be shown in the Table 18 as follows.

From the results displayed in the Table 18, the researcher found that 85.00% items of the 41 packages had the correlation point biserial > 0,30 or the discrimination index of these items belonged to the "Good" category. On the other hand, 15.00% items of the 41 packages had the correlation point biserial < 0.20 or the discrimination index of these items belonged to the "Not Good" category and, therefore, should be replaced. The average score of discrimination index from the 41 packages belonged to the "Good" category (the correlation point biserial > 0,30). The reason was that the values of the average correlation point biserial had been 0.472.

Based on the results displayed in the Table 19, the researcher found that 82.50% items of the package 42 had good discrimination index (correlation point biserial > 0,30), 2.50% items of the package 42 had quite good discrimination index (correlation point biserial ranged between 0.20 to 0.30) and 15.00% items of the package 42 did not have good discrimination index (correlation point biserial < 0.20). The average value of the correlation point biserial for the test packaged 42 had been equal to 0.513. This shows that the average discrimination index of 42 package are good categories (correlation point biserial values > 0,30).

Based on the results displayed in the Table 20, the researcher found that that 75.00% items of the package 43 had good discrimination index (correlation point biserial > 0.30) and 25.00% items of the package 43 did not have good discrimination index (correlation point biserial < 0.20). The average value of the correlation point biserial for the test packaged 43 had been equal to 0.440. These results showed that the average score of discrimination index for the packaged 43 belonged to the "Good" category (correlation point biserial values > 0.30).

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Table 18. 41 Package Discrimination Index

Criteria	Number of Items
Good	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 35, 36, 37, 38, 39, 40
Quite Good	-
Not Good	3, 6, 26, 27, 33, 34

Table 19. 42 Package Discrimination Index

Criteria	Number of Items
Good	1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 40
Quite Good	27
Not Good	2, 5, 6, 14, 34, 39

Table 20. 43 Package Discrimination Index

Criteria	Number of Items
Good	1, 3, 4, 5, 6, 7, 8, 7, 9, 10, 11, 12, 13, 14, 17, 18, 21, 22, 23, 25, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38
Quite Good	-
Not Good	2, 15, 16, 19, 20, 24, 28, 29, 39, 40

Table 21. 44 Package Discrimination Index

Criteria	Number of Items
Good	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 24, 25, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40
Quite Good	-
Not Good	5, 17, 18, 22, 23, 26, 27, 29, 38

Table 22. 45 Package Discrimination Index

Criteria	Number of Items
Good	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 40
Quite Good	39
Not Good	6, 17, 18, 20, 21, 22, 23, 32, 36

Table 23. 41 Package Difficulty Index

Criteria	Number of Items
Not good (Easy)	1, 35
Good (Medium)	2, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 36, 37, 38, 39, 40
Not good (Hard)	3, 6, 26, 27, 33, 34

Table 24. 42 Package Difficulty Index

Criteria	Number of Items
Not good (Easy)	11, 12, 16, 30
Good (Medium)	1, 3, 4, 7, 8, 9, 10, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 35, 36, 37, 38, 40
Not good (Hard)	2, 5, 6, 34, 39

Table 25. 43 Package Difficulty Index

Criteria	Number of Items
Not good (Easy)	16, 25
Good (Medium)	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 22, 23, 24, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38
Not good (Hard)	2, 15, 19, 20, 28, 29, 39, 40

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Table 26. 44 Package Difficulty Index

Criteria	Number of Items
Not good (Easy)	1, 2, 31, 36, 40
Good (Medium)	3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 28, 30, 32, 33, 34, 35, 37, 39
Not good (Hard)	5, 17, 23, 26, 27, 29, 38

Table 27. 45 Package Difficulty Index

Kategori	Nomor Butir Soal
Kurang (Mudah)	1, 39
Baik (Sedang)	2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 40
Kurang (Sukar)	6, 17, 18, 20, 21, 22, 23, 32, 36

Table 28. 41 Package Distractor Effectiveness

Criteria	Number of Items
Good	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 35, 36, 37, 38, 39, 40
Not good	3, 6, 26, 27, 33, 34

Table 29. 42 Package Distractor Effectiveness

Criteria	Number of Items
Good	1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 40
Not good	2, 5, 6, 14, 34, 39

Table 30. 43 Package Distractor Effectiveness

Criteria	Number of Items
Good	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 21, 22, 23, 25, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38
Not good	2, 15, 19, 20, 24, 28, 29, 39, 40

Tabel 31. 44 Package Distractor Effectiveness

Criteria	Number of Items
Good	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 24, 25, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40
Not good	5, 17, 18, 22, 23, 26, 27, 29, 38

Tabel 32. 45 Package Distractor Effectiveness

Criteria	Number of Items
Good	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 39, 40
Not good	6, 17, 18, 20, 21, 22, 23, 32, 36

Table 33. Model of Fit

Package	Number of Item		
	1PL	2PL	3PL
41	30	35	33
42	24	36	26
43	21	28	26
44	21	36	26
45	13	28	27

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Table 34. 41 Package Output *Bilogmg 3.0*

Parameter	Criteria	Number of Items
Discriminant Index (a)	Good	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40
	Not good	6, 34
Difficulty Index (b)	Good	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 35, 36, 37, 38, 39, 40
	Not good	3, 6, 26, 27, 33, 34

Table 35. 42 Package Output *Bilogmg 3.0*

Parameter	Criteria	Number of Items
Discriminant Index (a)	Good	1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
	Not good	6, 11, 24
Difficulty Index (b)	Good	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
	Not good	2, 5, 6, 11, 24, 34, 39

Table 36. 43 Package Output *Bilogmg 3.0*

Parameter	Criteria	Number of Items
Discriminant Index (a)	Good	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39
	Not good	1, 2, 19, 28, 40
Difficulty Index (b)	Good	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 22, 23, 25, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 38
	Not good	2, 15, 16, 19, 20, 24, 28, 29, 39, 40

Table 37. 44 Package Output *Bilogmg 3.0*

Parameter	Criteria	Number of Items
Discriminant Index (a)	Good	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
	Not good	1, 17, 18, 22, 26
Difficulty Index (b)	Good	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 24, 25, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40
	Not good	5, 17, 18, 22, 23, 26, 27, 29, 38

Tabel 38. 45 Package Output *Bilogmg 3.0*

Parameter	Criteria	Number of Items
Discriminant Index (a)	Good	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 39, 40
	Not good	17, 21, 22, 32, 36
Difficulty Index (b)	Good	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 39, 40
	Not good	6, 17, 18, 20, 21, 22, 23, 32, 36

The results displayed in the Table 21 showed that 77.50% items of the package 44 had correlation point biserial > 0.30 or the discrimination of the package 44 had been good. Meanwhile, 22.50% items of the package had correlation point biserial < 0.20 or the discrimination index of the package 44 had been not good. The discrimination index average scire of the package 44 belonged to the "Good" category (correlation point biserial values > 0.30). The reason was that the correlation point biserial score of the package 44 had been equal to 0.456.

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Based on the results displayed in the Table 22, the researcher found that 75.00% items of the package 45 had good discrimination index (correlation point biserial > 0.30), 2.50% items of the package 45 had quite good discrimination index (correlation point biserial ranged between 0.20 and 0.30) and 22.50% items of the package 45 had not good discrimination index (correlation point biserial < 0.20). The discrimination index average score of the package 45 belonged to the "Good" category (correlation point biserial values > 0.30). The reason was that the correlation point biserial average score of the package 45 had been equal to 0.441.

The difficulty index of each Junior High School Mathematics Examination test package might be seen from the proportion of correct answer. The difficulty index of each package would be shown in the following Table 23.

From the results displayed in the Table 23, the researcher found that 2.00% items of the package 41 belonged to the "Easy" category, 80.00% items of the package 41 belonged to the "Medium" category and 15.00% items of the package 41 belonged to the "Difficult" category. In other words, the researcher might conclude that 80.00% items had good difficulty index and 20.00% items had not good difficulty index. The test difficulty index average score of the package 41 belonged to the "Medium" and the "Good" category (the proportion of the correct answer ranged between 0.30 and 0.70). The reason was that the average proportion of the correct answer in the package 41 had been equal to 0.531.

From the results displayed in the Table 24, the researcher found that 10.00% items of the package 42 belonged to the "Easy" category (the proportion of correct answer > 0.70), 77.50% items of the package 42 belonged to the "Medium" category (the proportion of correct answer ranged between 0.30 and 0.70) and 12.50% items of the package 42 belonged to the "Difficult" category (the proportion of correct answer < 0.30). Then, 77.50 % items of the package 42 had good difficulty index and 22.50% items of the package 42 had not good difficulty index. The average value within the correct answer proportion of the package 42 had been equal to 0.562. The average value showed that the difficulty index average score of the package 42 belonged to the "Medium" and "Good" (the proportion of correct answer ranged between 0.30 and 0.70).

From the results displayed in the Table 25, the researcher found that 5.00% items of the package 43 belonged to the "Easy" category (the proportion of correct answer > 0.70), 75.00% items of the package 43 belonged to the "Medium" category (the proportion of correct answer ranged between 0.30 and 0.70) and 20.00% items of the package 43 belonged to the "Difficult" category (the proportion of correct answer < 0.30). In other words, 75.00 % items of the package 43 had good difficulty index, while 25.00% items of the package 43 had not good difficulty index. The average score in the correct answer proportion of the package 43 had been equal to 0.510. The average score showed that the average score of difficulty index for the package 42 had been equal to the "Medium" and "Good" category (the proportion of correct answer ranged between 0.30 and 0.70).

From the results displayed in the Table 26, the researcher found that 12.50% items of the package 44 belonged to the "Easy" category, 70.00% items of the package 44 belonged to the "Medium" category and 17.50 % items of the package 44 belonged to the "Difficult" category. In other words, 70.00% items had good difficulty index while the remaining 30.00% items had not good difficulty index. The average score of difficulty index for the package 44 belonged to the "Medium" and "Good" category (the proportion of correct answer ranged between 0.30 and 0.70). The reason was that the average proportion of correct answer for the package 44 had been equal to 0.522.

From the results displayed in the Table 27, the researcher found that 5.00% test items of the package 45 belonged to the "Easy" category (the proportion of correct answer > 0.70), 72.50% test items of the package 45 belonged to the "Medium" category (the proportion of correct answer ranged between 0.30 and 0.70) and 22.50% test items of the package 45 belonged to the "Difficult" category (the proportion of correct answer < 0.30). in other words, 72.50% test items of the package 45 had good difficulty index and the remaining 27.50% test items of the package 45 had not good difficulty index. The average score of difficulty index for the package 45 belonged to the "Medium" and "Good" category (the proportion of correct answer ranged between 0.30 and 0.70). The reason was that the average score in the proportion of correct answer for the package 45 had been equal to 0.504.

Print ISSN: 2356-2684, Online ISSN: 2477-1503

The distractor effectiveness might be seen from the number of test participants who chose the distractor and the correlation value of each point biserial distractors in the output generated by the Microcat Iteman software.

The results displayed in the Table 28 showed that 85.00% items of the package 41 had good distractor effectiveness because each distractor had been chosen at least by 5% of the participants test and had a negative biserial correlation point. On the other hand, 15% items of the package 41 did not have distractor effectiveness because one of the distractors had been selected by less than 5% of participants test or had a positive correlation point biserial.

Based on the results displayed in the Table 29, 85.00% items of the package 42 had good distractor effectiveness because each distractor had been chosen at least by 5% of the participants test and had a negative biserial correlation point. On the contrary, 15.00% items of the package 41 did not have good distractor effectiveness because one of the distractors had been selected by less than 5% of the participants test or had a positive correlation point biserial.

Based on the results displayed in the Table 30, 77.50% items of the package 43 had good distractor effectiveness because each distractor had been chosen at least by 5% of participants test and had a negative biserial correlation point. On the contrary, the remaining 22.50% items of the package 43 did not have not good distractor effectiveness because one of the distractors had been selected by less than 5% of participants test or have a positive correlation point biserial.

Based on the results displayed in the Table 31, 77.50% items of the package 44 had good distractor effectiveness because each distractor had been chosen at least by 5% of the participants test and had a negative biserial correlation point. While 22,5% items of 44 packages have not good distractor effectiveness because one of the distractors have less than 5% of participants test or have a positive correlation point biserial.

Based on Table 32, 77,5% items of 45 packages have good distractor effectiveness because each distractor was chosen at least 5% of participants test and has a negative biserial correlation point. Meanwhile, 22.50% items of the package 45 did not have good distractor effectiveness because one of the distractors had been selected by less than 5% of participants test or had a positive correlation point biserial.

The Quality of Tests are Based on Item Response Theory

For the quantitative analysis by means of item response theory, the researcher ran the Bilogmg 3.0 software. In order to determine the most suitable logistics parameters, it was necessary to test the mode suitability. The model compatibility might be determined by implementing the chi squared for each parameter logistic of the model. The chi squared table for each test package would be presented as follows.

Based on the results displayed in the Table 33, the researcher found that all packages of Junior High School Mathematics Examination in Regency of Bangkalan area had been fit into the 2PL model. This was seen from the number of test items that had the most suitable model than the 1PL model and the 3PL model. The parameters that should be estimated were the discrimination index (a) and the difficulty index (b).

The output of Bilogmg 3.0 that had been implemented for the analysis was the output phase 2, which had been the output that contained the parameter estimation of these items. The output results might be summarized in the following Table 34.

First, based on the results displayed in the Table 34 the researcher found that: (1) there had been 95.00% items of the package 41 which had good discrimination index ($0.00 \leq a \leq 2.00$); (2) there had been 85.00% items that had good difficulty index ($-2.00 \leq b \leq 2.00$); and (3) there had been two items that could not be analyzed, namely the item number 6 and number 34.

Second, based on the results displayed in the Table 35 the researcher found that: (1) there had been 92.50% items of the package 42 which had good discrimination index ($0.00 \leq a \leq 2.00$); (2) there had been 82.50% items that had good difficulty index ($-2.00 \leq b \leq 2.00$) and (3) there had been 3 items that could not be analyzed namely the item number 6, number 11 and number 24.

Third, based on the results displayed in the Table 36 the researcher found that: (1) there had been 87.50% items of the package 43 that had good discrimination index ($0.00 \leq a \leq 2.00$); (2) there had been 75.00% items that had good difficulty index ($-2.00 \leq b \leq 2.00$); and (3) there had been 4 items that could not be analyzed namely the item number 2, number 19, number 28 and number 40.

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Fourth, based on the results displayed in the Table 37 the researcher found that: (1) there had been 87.50% items of the package 44 that had good discrimination index ($0.00 \leq a \leq 2.00$); (2) there had been 77.50% items that had good difficulty index ($-2.00 \leq b \leq 2.00$); and (3) there had been 4 items that could not be analyzed namely the item number 17, number 18, number 22 and number 26.

Based on Table 38, it can be seen that: (1) there is 87,5% items of 45 packages which have good discrimination ($0 \leq a \leq 2$); (2) there is 77,5% items that have good difficulty index ($-2 \leq b \leq 2$) and (3) there are 5 items that can not be analyzed, they are number 17, 21, 22, 32, and 36.

Test Equating

For the analysis Analysis of Junior High School Mathematics Examination test equating in the Regency of Bangkalan area, the researcher implemented the test characteristic curve method. Test characteristic curves of each package might be combined in order to determine the test equivalence. The result of merging the test characteristic curve methods might be seen in the following figure.

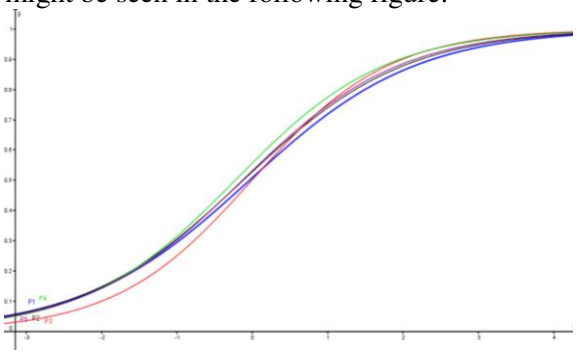

Figure 2. Test Characteristic Curve

Discussion

Tests Quality

The qualitative analysis toward the test plan of Junior High School Mathematics Examination in Regency of Bangkalan area produced the "Quite Good" category or the test plan might be implemented with minor revisions. The test plan should be improved on the indicators aspects that were too specific and on the suitable indicators with basic competence because the selection of the operational verb had been less proper due to the form or the appearance of the test plan. From the qualitative analysis toward the lattice test of Junior High School Mathematics Examination in the Regency of Bangkalan area, the researcher would like to conclude that the test plan that would be used for the test would consist of some packets and these packets should be made only under one test plan category which had been suitable for the mathematics competence standards of the Junior High School graduates.

From the qualitative analysis of the package 41 in the Junior High School Mathematics Examination within the Regency of Bangkalan area, there had been 45.00% test items that did not have good quality. Meanwhile, 55.00% test items had quite good quality with minor revisions. Within the package 42, there had been 12.50% test items that had good quality in terms of material, construction and language aspect. Then, there had been 40.00% test items that did not have good quality. Next, 47.50% test items had quite good quality with minor revisions. Furthermore, within the package 43 there had been 2.50% test items that had good quality. Then, there had been 27.50% test items that did not have good quality. Last but not the least, 70% test items had quite good quality with minor revisions. In the package 44 of the Junior High School Mathematics Examination for the Regency of Bangkalan Area there had been 2.50% items that had good quality. Then, 55.00% test items that did not have good quality. Next, 42.50% test items had quite good quality with minor revisions. Eventually, within the package 45 there had been 62.50% test items that did not have good quality. Meanwhile, the remaining 37.50% items had quite good quality with minor revisions.

Some test items had been categorized as "Not Good" because these items had not been in accordance with the indicators on the test plan, had not been suitable for measuring the achievement and the logical of test item alternatives. Some test items had fallen into the "Quite Good" category and should be given minor revisions. Within the improvements toward the construction aspect, some multiple choice alternatives that took the form of numbers had not been sorted yet and had been lack of clarity of images or graphics. In the aspect of language, there should be improvement in the language, the punctuation and the grammar of some test items and there was less communicative language that had been used.

Based on the quantitative analysis by means of classical test theory approach, all of the test packages in the Junior High School Mathematics Examination for the Regency of

Print ISSN: 2356-2684, Online ISSN: 2477-1503

Bangkalan area had good quality. All of the packages had good content validity evidence (0.90, 0.94, 0.93, 0.81 and 0.83) and good criteria validity (0.525, 0.528, 0.555, 0.485 and 0.496). All of these packages had high reliability estimation (0.917, 0.921, 0.907, 0.913 and 0.910). All of the test packages had good item difficulty index ($\overline{p}$ = 0.531, 0.526, 0.510, 0.521 and 0.504). All of these packages also had good discriminant index ($\overline{rpbis}$ = 0.472, 0.513, 0.440, 0.456 and 0.442). Last but not the least, al of these packages had good distractor effectiveness (85.00%, 85.00%, 77.50%, 77.50% and 77.50%).

The results of the analysis by means of item response theory showed that all of the test packages had possessed good quality. All of the discriminant index package belonged to the "Good" category (0.887, 0.911, 0.945, 1.006 and 0.952). The difficult index from all of the test packages also belonged to the "Good" category (0.023, 0.113, 0.045, 0.215 and 0.113). Within the test packages there were some tests items that belonged to the "Difficult" category. This was due to several factors, for instance, the unclear, the unclear graphs, the unclear tables, the unclear diagram, or the existing issues on test items that had not been in accordance to the rules.

Based on the qualitative and quantitative analysis, the quality of Junior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan area had not been in the "Good" category. In the package 41, 40.00% test items had quite good quality and 60.00% test items had not good quality.Then, in the package 10.00% test items had good quality, 32.50% tests items had quite good quality and 57.50% test items had not good quality. Next, in the package 43 2.50% test items had good quality, 47.50% test items had quite good quality and 50.00% test items had not good quality. Furthermore, in the package 44 32.50% had quite good quality and 67.50% had not good quality. Last but not the least, in the package 45 27.50% test items had quite good quality and 72.50% test items had not good quality.

Tests Equating

The Junior High School Mathematics Examination in the Regency of Bangkalan area might be considered equal. The equality might be seen from the test characteristic curves of each package that had been adjacent or that had been nearly coinciding. The equation of the Junior High School Mathematics Examination test in the Regency of Bangkalan area made use of the curve characteristic method. The equalization within curve characteristic method was influenced by the difficulty index and the discrimination index. Both the difficulty index and the discrimination index belonged to the same categories.

CONCLUSIONS

The test plans quality of Junior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan area belongs to the "Quite Good" categories. However, the quality of the Juniior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan area belongs to the "Not Good" category. In the package 41, 40.00% test items have the quite good quality and 60.00% test items have the not good quality. Then, in the package 42 10.00% test items have good quality, 32.50% test items have quite good quality and 57.50% test items have not good quality. Next, in the package 43, 2.50% test items have good quality, 47.50% test items have quite good quality and 50.00% tests items have not good quality. Furthermore, in the package 44, 32.50% test items have quite good quality and 67.50% test item have not good quality. Last but not the least, 27.50% test items have quite good quality and 72.50% test items have not good quality.

Based on the characteristic curve test method, the Junior High School Mathematics Examination for the 2015/2016 Academic Year in the Regency of Bangkalan area has been considered equal. The equality might be seen from the characteristics curve test toward each package that has been adjacent or that has been nearly coinciding.

REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*. *Vol. 40, 955-959.*

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, California USA: Brooks/Cole Publishing Company.

Ary, D., Jacobs, L.C., Sorensen, C., & Razavieh, A. (2010). *Introduction of research in*

Print ISSN: 2356-2684, Online ISSN: 2477-1503

education 8th edition. Belmont, California USA: Wadsworth.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: the position of correct answers in multiple-choice test items items as a psychometric variable. *Journal of Educational Measurement, 40, 2, 109-128.*

Budiman, A., & Jailani, J. (2014). Pengembangan instrumen asesmen higher order thinking skill (HOTS) pada mata pelajaran matematika junior high school kelas VIII semester 1. *Jurnal Riset Pendidikan Matematika 1*(2). 139-151. doi:http://dx.doi.org/10.21831/jrpm.v1i2.2671

Chauhan, P., et al. (2015). Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *International Journal of Anatomy and Research, Vol. 3, No. 4, 1607-1610.*

DeMars, C. (2010). *Item response theory.* Oxford, New York USA: Oxford University Press, Inc.

Ebel, R. L & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliffs, New Jersey USA: Prentice-Hall, Inc.

Faremi, Y. A. (2016). Reliability coefficient of multiplechoice and short answer objective test items in basic technology: comparative approach. *Journal of Educational Policy and Entrepreneurial Research (JEPER), 3, 3, 59-69.*

Gunartha, I W., Kartowagiran, B., & Suardiman, S. (2014). Pengembangan model evaluasi program layanan pendidikan anak usia dini (PAUD). *Jurnal Penelitian dan Evaluasi Pendidikan, 18*(1), 30-43. doi:http://dx.doi.org/10.21831/pep.v18i1.2122

Jailani, J., & Retnawati, H. (2016). The challenges of junior high school mathematic teachers in implementing the problem-based learning for improving the higher-order thinking skills. *The Online Journal of Counseling and Education, 5*(3), 1-13. Retrived from http://www.tojce.com/volume/volume-5-issue-3

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement,* (*7th edition*). Hoboken, New Jersey USA: John Wiley & Sons, Inc.

Mardapi, D. (2005). *Pengembangan instrumen penelitian pendidikan*. Yogyakarta: Nuha Litera.

Mardapi, D. (2012). *Pengukuran penilaian & evaluasi pendidikan.* Yogyakarta: Nuha Litera.

Miller, M.D., Linn, R.L., & Gronlund, N.E. (2009). *Measurement and assessment in teaching*. Upper Saddle River, New Jersey USA: Pearson.

NCTM. (2000). *Principles and standards for school mathematics.* Reston, Virginia USA: The National Council of Teachers of Mathematics, Inc.

Nitko, A. J. & Brookhart, S. M. (2007). *Educational assessment of students*. Englewood Cliffs. New Jersey USA: Prentice-Hall, Inc.

Nurlita, M. (2015). Pengembangan soal terbuka (open-ended problem) pada mata pelajaran matematika SMP kelas VIII.*PYTHAGORAS: Jurnal Pendidikan Matematika, 10*(1), 38-49. doi:http://dx.doi.org/10.21831/pg.v10i1.9106

Popham, W. J. (2009). *Instruction that measures up. Succsessful teaching in the age of accountability*. Alexandria, Virginia USA: ASCD.

Retnawati, H. (2013). *Teori respons butir dan penerapannya*. Parama: Yogyakarta

Retnawati, H. (2015). The equating of the test of english proficiency (TOEP). *International Conference on Education, Psychology and Society (ICEEPS)-1801, 276-287.*

Retnawati, H. (2016). Analisis kuantitatif instrumen penelitian. Parama: Yogyakarta

Salvia, J., Ysseldyke, J. E., Bolt, S. (2010). *Assessment in special and inclusive education.* Belmont, CA USA:Wadsworth.

Urbina, S. (2014). *Essentials of psychological testing (2nd edition).* Haboken,NJ USA:John Wiley & Sons, Inc.

Von Davier, A. A. (2011). *Statistical models for test equating, scaling and linking.* Princeton, New Jersey USA: Springer.

Print ISSN: 2356-2684, Online ISSN: 2477-1503