CHAPTER IV

DESCRIPTION, RESEARCH FINDING AND DATA ANALYSIS

Chapter IV presents description of the data, research finding which containts analysis of validity, reliability, practicality and item analysis of English summative test in term of a good test, analysis of the data, discussion and limitation of the research.

A. Description of the Data

In this chapter, the writer had analyzed the data gathered from the research. The data was obtained from: 1.) The students' answer sheet and English summative test question paper for the second grade of MA Uswatun Hasanah Semarang made by MGMP LP Ma'arif NU of Semarang district at odd semester in the academic year 2013/2014. The writer used English summative test which was conducted on Saturday, December 7th 2013, from 07.30 to 09.00 a.m. The test consists of 55 items, 50 multiple choice items and 5 essay items. 2.) The result of interview to the teacher is used to know the practicality of the test. 3.) The English syllabus.

Having gained the whole needed data, the researcher then did analysis which refers to the statistical data analysis to find out whether or not the test is categorized as a good test. As Robert L. Thorndike and Elizabeth Hagen stated that a well constructed test should have three main characteristics which involve validity, reliability, and practicality.[85] The test is called valid if it can measure what is supposed to measure, it can be reliable if the result of the test is the same even though the test administered to the same level students in the next time and it can be practical if it is easy to do and administer. The test also should good at its item analysis to be a good test. The item analysis contains three components, they are difficulty level, discriminating power and the effectiveness of the distractor.

[85] Sudijono, Anas, *Pengantar Evaluasi Pendidikan*, (Jakarta: PT Raja Grafindo Persada, 2008),p.93

According to Purwanto, good items test is test which have moderate difficulty level, high discriminating power and distractor analysis which work effectively.[86]

The first thing to do before conducting analysis was classifying the students based on the test result. The researcher divided 39 students into three classification they are upper group, middle group and lower group. Here, only two groups that is used to analyze, they are upper and lower group. Meanwhile, the middle group is disregarded. The table below is the students position in the group based on the test result.

Table 4

The group position based on the test result

No.	Students	Score	Group
1.	Code Student 14	82	Upper Group
2.	Code Student 4	74	Upper Group
3.	Code Student 11	71	Upper Group
4.	Code Student 13	71	Upper Group
5.	Code Student 16	70	Upper Group
6.	Code Student 18	70	Upper Group
7.	Code Student 12	69	Upper Group
8.	Code Student 17	67	Upper Group
9.	Code Student 20	67	Upper Group
10.	Code Student 19	65	Upper Group
11.	Code Student 15	64	Upper Group
12.	Code Student 35	61	Upper Group
13.	Code Student 7	58	Upper Group
14.	Code Student 9	57	Middle Group

[86] Purwanto, E*valuasi Hasil Belajar*, (Yogyakarta: Pustaka Pelajar, 2009), p. 99

15.	Code Student 10	55	Middle Group
16.	Code Student 32	55	Middle Group
17.	Code Student 31	54	Middle Group
18.	Code Student 3	52	Middle Group
19.	Code Student 6	52	Middle Group
20.	Code Student 22	49	Middle Group
21.	Code Student 23	49	Middle Group
22.	Code Student 24	49	Middle Group
23.	Code Student 38	49	Middle Group
24.	Code Student 25	48	Middle Group
25.	Code Student 30	47	Middle Group
26.	Code Student 8	47	Middle Group
27.	Code Student 2	46	Lower Group
28.	Code Student 34	46	Lower Group
29.	Code Student 36	46	Lower Group
30.	Code Student 39	46	Lower Group
31.	Code Student 33	45	Lower Group
32.	Code Student 1	44	Lower Group
33.	Code Student 28	44	Lower Group
34.	Code Student 37	42	Lower Group
35.	Code Student 26	41	Lower Group
36.	Code Student 5	41	Lower Group
37.	Code Student 27	40	Lower Group
38.	Code Student 29	40	Lower Group
39.	Code Student 21	39	Lower Group

B. Research Finding

1. Content Validity of the Test

Doing analysis of content validity, the researcher did by comparing between the contents of the test and the indicator that recommended in the syllabus. The indicators indicate which are found in the items of English summative test is divided by the indicators which are recommended in the syllabus. Then, the result is multiplied in to 100%. The result of comparing is called the conformity level.

As Suharsimi Arikunto states that the formula of conformity level as follow[87]:

$$C = \frac{A}{B} \; x\; 100\ \%$$

Explanation :

C : Conformity level

a : The indicators indicate which are found in the items of English summative test.

b : The indicators which are recommended in the syllabus.

The table below is the result of content validity analysis. The detail analysis can be seen in appendix:

[87] Arikunto, Suharsimi, *Prosedur Penelitian*, (Jakarta: Rhineka Cipta, 2002), p.313

Table 5

The result of content validity analysis

No	Skill	Item Number	Total	Indicator found in the items of English summative test	Total indicator recommended in the syllabus
1.	Reading	2, 3, 4, 15, 16, 17, 18, 21, 22, 23, 31, 37, 28, 40, 9, 10, 41, 45, 46, 35, 39, 44, 24, 32, 29, 36	27 items	7 indicators indicated	
2.	Writing	2, 3, 4, 9, 10, 13, 15, 16, 17, 18, 39, (1, 2 and 3 essay), 28, 31, 47, 21, 22, 23, 32, 33, 34, 35, 36, 41, 44, 29	28 items	7 indicators indicated	
Total				14 indicators indicated	23 indicators

From the data above, there are 14 indicators indicate which are found in the items of English summative test distributed in to 50 items in 2 skills. Next, There are 23 indicators indicate which are recommended in the syllabus. Then, the researcher calculated the data to know the validity as follow:

$$C = \frac{14}{23} \text{ x } 100\%$$

$$C = 0{,}608 \text{ x } 100\%$$

$$C = 60{,}8\%$$

Finishing calculate the data, then, the researcher concluded that the content validity of the test is considered as adequate in term of content validity, because it is range from 56% to 75%. It means that the contents of the test have covered adequately to the indicators which are recommended in the syllabus.

The contents of the test did not cover all indicators which are recommended in the syllabus. All of indicators which did not found in the contents of the test are about identifying the texts especially in identifying the generic structure of the texts. There were 9 indicators did not find in the contents of the test, 6 indicators are about identifying the texts especially in identifying the generic structure of the texts. Most of contents of the test just focused on the finding the meaning of the texts and did not contained indicators of identifying the generic structure of text. It is make all of the contents of the test did not cover the indicators. In the other hand, one of the principles in constructing the test are sample of the test items should be representative with the teaching material which have been stated in the syllabus. In this case, that principle did not be applied in the test.

2. Reliability

Reliability refers to the consistency of evaluation results. If different teacher independently rate the same students in the same instrument and obtain similar ratings, we can conclude that the result has a high degree of reliability from one rate to another.

a) Reliability of Multiple Choice Test

To know the realiability of multiple choice test, the researcher used *K-R 20* formula.[88]

$$r_{11} = \frac{n}{n-1}\left(\frac{S_t{}^2 - \sum p_i q_i}{S_t{}^2}\right)$$

Explanation :

r_{11} : coefficient of reliability tests

n : the number of testee who joined in test

1 : constant number

S_t : total variance

p_i : proportion of the testee who answered the items correctly

q_i : proportion of the testee who answered the items incorrectly

$p_i\, q_i$: the result number of multiplication $p_i\, q_i$

in which, Finding S_t^2 variants using the formula:

$$S_t{}^2 = \frac{\sum X_t{}^2 - \left(\frac{\sum X_t}{N}\right)^2}{N}$$

Explanation :

N : the number of test items

X_t : total score of test item

Then, the researcher calculated the reliability for multiple choice test. The first computation is finding $S_t{}^2$:

[88] Purwanto, *Evaluasi Hasil Belajar,* (Yogyakarta: Pustaka Pelajar, 2009), p. 170-171

$$S_t^{\ 2} = \frac{9456 - \left(\frac{474}{50}\right)^2}{50}$$

$$S_t^{\ 2} = \frac{9456 - 89.87}{50}$$

$$S_t^{\ 2} = 187.322$$

After finding $S_t^{\ 2}$, the computation is continued to compute r_{11} :

$$r_{11} = \frac{26}{26-1}\left(\frac{187 - 9.020}{187.322}\right)$$

$$= \frac{26}{25}\left(\frac{177.98}{187.322}\right)$$

$$= 0.989$$

From the computation above, the result of computing reliability of the multiple choice test is 0.989. Then, the researcher interpreted the coefficient reliability of the test according to Anas Sudijono stated that, If r_{11} is equal or more than 0.70, it means that the test results of the study have high reliabillity.[89] Since the result of computation is higher than 0.70, the researcher concluded that the multiple choice test is considered as reliable.

b) Reliability of Essay Test

Doing evaluating the reliability of essay test, the researcher used the formula *alpha*:[90]

$$r_{11} = \frac{k}{k-1}\left(1 - \frac{\sum Si^2}{St^2}\right)$$

[89] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.209

[90] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.208

Explanation:

r_{11} : coefisien reliability of the test

k : total item

1 : constant's number

Si : varians score of each item

St : varians total

The researcher calculated the reliability for essay test. Here is the result:

$$r_{11} = \frac{5}{5-1}\left(1 - \frac{22.87}{64.42}\right)$$

$$= \frac{5}{4}(1 - 0.355)$$

$$= 1.25(0.6445)$$

$$= 0.806$$

After the researcher calculated the data, the researcher found that the result of calculating the reliability of essay test is 0.806. It is similar with the multiple choice test that the essay test also would be interpreted to the coefficient reliability of the test. Thus, Since the result of computation is higher than 0.70, the researcher concluded that the essay test is also considered as reliable.

At both point multiple choice and essay test similarly has high of reliability. Multiple choice has 0.989 and essay test has 0.806 of reliability. As what have been explained in the chapter II, the fifth principle of constructing the test that the should be high reliability. So that, the test will have some result even it is given to the same student on two different occasions, it will produce similar results.

3. Practicality of the Test

In selecting a test and other instruments, practical considerations cannot be neglected. One of the factors relevant to the practicality when selecting tests is ease of administration. In fact, ease of administration involves the simple and clear directions, the subtests in minimum numbers and the easy timing. As Anas Sudjijono stated that, a good practicality on the test means that the test should have two criteria they are simple and complete.[91] Then, the researcher did analysis of practicality of the test by doing interview technique. The researcher interviewed to the teacher to get some data about the practicality of the English summative test that has been mentioned before. The following table is the result of analysis from the interview.

Table 6

The analysis result of practicality

No.	Criteria of good practicality		Yes	No
1.	*Simple*	No need much tools to do and difficult tools to find	✓	
2.	*Complete*	The instructions how to do the test	✓	
		The answer key	✓	
		The scoring guidance		✓

From the table above practicality has two criteria. They are simple and complete. Simple means that it does not require much equipment or tools that are difficult to get and complete means that the test comes completed by the instructions on how to do it, the answer keys and the scoring guidance to guide the teacher to score the test.

[91] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan*, (Jakarta: PT Raja Grafindo Persada, 2008), p.97

The practicality requirements of the test according to the table above, The test had the requirement of simple that is no need much equipments to do the test and no difficult to find the tools. It was proved by the teacher statement that, “the students can use their pencil or ballpoint pen to do the test and if they donot have these, they can buy in our school cooperation”.Here, the tools used to do the test was so simple they are pencill or ballpoint pen and these tools was supported by the axisted of school cooperation in their school.

Meanwhile, for the complete requirements, the test is also completed by the instruction on how to do it. It is proved in the English summative test question paper. There are many instructions to do the test not only in multiple choices but also in essay test. For example, “choose the right answer by crossing *a, b, c, d,* or *e*!” and “arrange these words into correct sentences and put the correct tobe!”. The test also completed with the key answer given by Lembaga Pendidikan Ma’arif NU of Semarang district but, not for the scoring guidance. It made the teacher confused in scoring the test and gave the score as he wanted it.

4. Difficulty Level of the Test

a. Difficulty Level of Multiple Choice Test

The following is the computation of difficulty level for multiple choice item test number 1 and for the other items would use the same formula.

$$P = \frac{B}{JS}$$

Explanation :

P : Index of difficulty

B : The total number of students who got the item correct

JS : The number of students who took a test[92]

Then, the researcher calculated the data as follow:

$B = 9$ $\qquad$ $JS = 26$

$$P = \frac{B}{JS}$$

$$P = \frac{9}{26} \qquad P = 0.346$$

It is proper to say that the index of difficulty of the item number 1 above can be said as the moderate category, because the calculation result of the item number 1 is in the interval $0.30 < p < 0.70$.

After computing 50 items of the multiple choice test, there are 4 items are considered as easy, 25 items are considered as moderate and 21 items are considered as difficult. The whole computation result of difficulty level can be seen in appendix. The following table is the result of analyzing difficulty level of multiple choice test.

Table 7

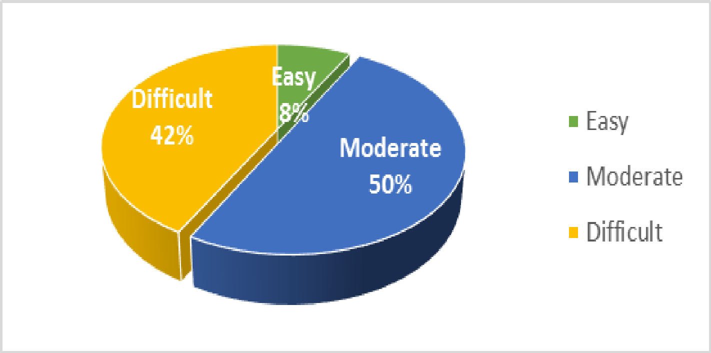
Difficulty level of multiple choice test

No.	Criteria	Item Number	Total item	Percentage
1.	*Easy*	21, 22, 23, 24.	4 items	8%
2.	*Moderate*	1, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 20, 28, 30, 31, 32, 34, 35, 38, 39, 40, 42, 43, 45, 50.	25 items	50%
3.	*Difficult*	2, 7, 8, 15, 16, 17, 18, 19, 25, 26, 27, 29, 33, 36, 37, 41, 44, 46, 47, 48, 49.	21 items	42%
Total			50 items	100%

Pie chart 1

[92] Arikunto, Suharsimi, *Dasar-dasar Evaluasi Pendidikan*, (Jakarta: PT. Bumi Aksara, 2005), edisi revisi, p. 208

Pie chart of the Difficulty Level in Multiple Choice test



The difficulty level in the multiple choice test items have only 50% which place in the rank of moderate level. Doing further analysis, the researcher found that it might be caused by there were some items that did not match with the basic rules of constructing multiple choice test. There were 17 questions (stem) of 50 multiple choice questions in the test which is so wordy. It would make the students booring and confusing in reading the questions. There was negative stem in the test, the test did not avoid negatively worded stems and it did not present a single clearly formulated problem in the stem of the item, it would also make the students feel difficulty in doing the test. The whole analysis about constructing the multiple choice test items can be seen in appendix.

b. Difficulty Level of Essay Test

After analyzing the difficulty level of multiple choice test. Then, the researcher analyzed the difficulty level of essay test. To know the reliability of essay test, the researcher used the formula as follow: [93]

$$\textit{Mean} = \frac{\textit{The total of Students score for each item}}{\textit{The number of Students}}$$

$$\textit{Index of Difficulty} \quad \frac{\textit{Mean}}{}$$

[93] Arifin, Zainal, *Evaluasi Pembelajaran*,(Bandung: Remadja Rosdakarya,2009), p.135

=

Maximum score of each item

The following is the computation of difficulty level for essay item test number 1 and for the other items would use the same formula.

Total of students' score for each item = 178

The number of students = 26

$$Mean = \frac{\textit{The total of Students score for each item}}{\textit{The number of Students}}$$

$$Mean = \frac{178}{26}$$

$$Mean = 6.846$$

After computing Mean, the calculation was continued to find the index of difficulty.

$$Index\ of\ Difficulty = \frac{Mean}{\textit{Maximum score of each item}}$$

$$Index\ of\ Difficulty = \frac{6.846}{10}$$

$$Index\ of\ Difficulty = 0.6846$$

The obtained result states that index of difficulty for essay test is *0.685* and after being consulted to the rank scale of difficulty level, it is found that the result is on 0.30< p < 0.70. thus, the item number 1 is considered as moderate. Here, the results of the difficulty level of essay test.

Table 8

Difficulty Level of Essay test

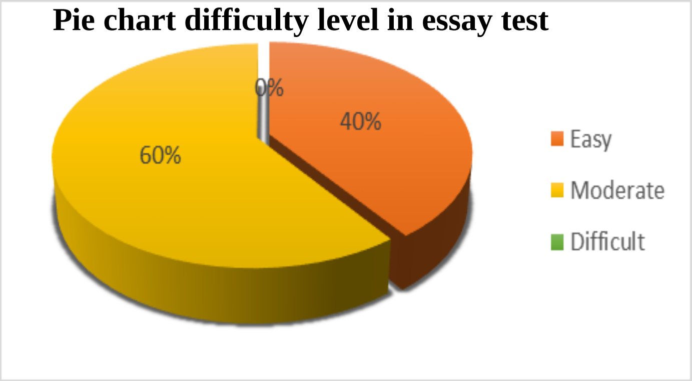
No. Item	Total of Students' Score	Mean	Index of Difficulty Level	Interpretation
51	178	6,846153846	0,684615385	*Moderate*
52	196	7,538461538	0,753846154	*Easy*
53	194	7,461538462	0,746153846	*Easy*
54	173	6,653846154	0,665384615	*Moderate*
55	136	5,230769231	0,523076923	*Moderate*

From the table above, the researcher found that the item number 51, 54 and 55 are categorized to be moderate. Meanwhile, the item number 52 and 53 are categorized to be easy. Then, the researcher made the percentage as follow:

Table 9
The difficulty level interpretation of essay test

No.	Interpretation	Item Number	Total Item	Percentage
1.	*Easy*	52 and 53	2 items	40%
2.	*Moderate*	51, 54 and 55	3 items	60%
3.	*Difficult*	-	0 item	0%
Total			5 items	100%

Pie chart 2



Finding that the essay test has 60% of difficulty index, then, the researcher did further analysis by analyzing the items of the test and the basic principles of constructing the essay test. Here, the researcher found that the test maker made the essay test contrast with the basic principles of constructing the essay test. One of the principles are the item test should represent the material but, the items test here did not represent all of material that have been taught. Test items which were categorized to be easy were the items test about using grammar accurately.

5. Discriminating Power of the Test

a. Discriminating Power of Multiple Choice Test

The discriminating power measures how well the test items arranged to identify the differences in the students' competence. To do this analysis, the researcher used the formula: [94]

$D = PA - PB$

In which, computing the *PA* and *PB* by using the formula:

$$PA = \frac{BA}{JA}$$

$$PB = \frac{BB}{JB}$$

Explanation:

D	: Discriminating power
PA	: Proportion of high group who answered the item correctly
PB	: Proportion of lower group who answered the item correctly
BA	: Total students in high group who answered the item correctly
JA	: Total students in high group

[94] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p. 389

BB : Total students in lower group who answered the item correctly

JB : Total students in lower group.

Then, the researcher calculated the discriminating power for multiple choice test. The following is the computation of the discriminating power for multiple choice test number 1 and for the other items would use the same formula.

$BA = 8 \qquad JA = 13$

$BB = 1 \qquad JB = 13$

$$PA = \frac{BA}{JA} \qquad PB = \frac{BB}{JB}$$

$$PA = \frac{8}{13} \qquad PB = \frac{1}{13}$$

$$PA = 0.615 \qquad PB = 0.077$$

After finishing compute the *PA* and *PB*, then, the researcher calculated the discriminating power :

$D = PA - PB$

$D = 0.615 - 0.077$

$D = 0.538$

The obtained result stated that the discriminating power of the item number 1 above can be said as a good category, because the calculation result of the item number 1 is in the interval $0.40< D <0.70$. The whole computation result of difficulty level can be seen in appendix.

After computing 50 items of the multiple choice test, the researcher found the result of discriminating power analysis that 14 items are considered as bad category, 13 items are considered as poor category, 6 items are considered as satisfactory category, 16 items are considered as good category and 1 item is considered as high category. The following table is the results of the discrimination index of the test items.

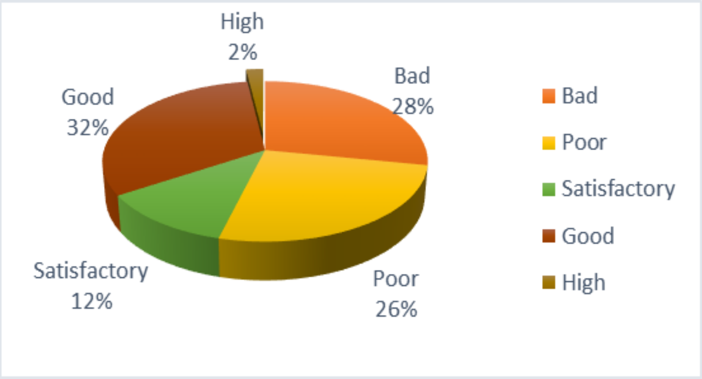
Table 10

The discrimination index of multiple choice test

No.	Category	Item Number	Total item	Percentage
1	*Bad*	2, 9, 10, 16, 19, 20, 26, 27, 29, 33, 41, 46, 47, 49.	14 items	28%
2	*Poor*	7, 12, 15, 17, 18, 23, 24, 25, 36, 37, 43, 44, 48.	13 items	26%
3	*Satisfactory*	8, 13, 14, 21, 22, 28, 31, 50.	6 items	12%
4	*Good*	1, 3, 4, 5, 6, 30, 32, 34, 35, 38, 39, 40, 42, 45.	16 items	32%
5	*High*	11.	1 item	2%
Total			50 items	100%

Pie chart 3

Pie chart of discrimination index in multiple choice test



To be a good discriminating power, the test should have high discriminating power. Here, the researcher found that the discriminating power of multiple choice items test had only 2% or 1 item. Doing further analysis, the researcher found that it could be happen because the items of the test had options of the correct answer which did not homogenized. There were some options were not plausible and were not logical. It would cause bad discrimiation items. The whole analysis can be seen in appendix.

b. Discriminating Power of Essay Test

To compute the essay items, the reseracher used the technique to analyze by using the formula: [95]

$$DP = \frac{\bar{x}\,ka\ -\ \bar{x}\,kb}{Max\ score}$$

Explanation :

DP : Discriminating power

$\bar{x}\,ka$: Average of upper group

$\bar{x}\,kb$: Average of lower group

Max score : Maximum score

The following was the example computation of discriminating power for essay item test number 1 and for the other items would use the same formula.

$\bar{x}\,ka$ = 7.692 $\bar{x}\,kb$ = 6

Maximum score = 10

$$DP = \frac{\bar{x}\,ka\ -\ \bar{x}\,kb}{Max\ score}$$

[95] Arifin, Zainal, *Evaluasi Pembelajaran*, (Bandung: PT Remaja Rosdakarya, 2011), p.133

$$DP = \frac{7.692 - 6}{10}$$

$DP = 0.169$

The obtained result states that index of discriminating power for essay test is *0.169* and after being consulted to the rank scale of discriminating power, it is found that the result is on 0.00< *D* <0.20. thus, the item number 1 is considered as poor category. The table below is the results of the difficulty level analysis of essay test.

Table 11

The Discrimination Index of Essay Test

No.	Average of Upper Group	Average of Lower Group	Index of Discriminating Power	Interpretation
51	7,692307692	6	0,169230769	*Poor*
52	9,538461538	5,53846154	0,4	*Satisfactory*
53	7,846153846	7,07692308	0,076923077	*Poor*
54	8,769230769	4,53846154	0,423076923	*Good*
55	7,076923077	3,38461538	0,369230769	*Satisfactory*

Based on the table above, the researcher concluded that the item number 51 and 53 are categorized to be poor, item number 54 is categorized to be good and item number 52 and 55 are categorized to be satisfactory. Then, the researcher made the percentage of the result calculation as follow

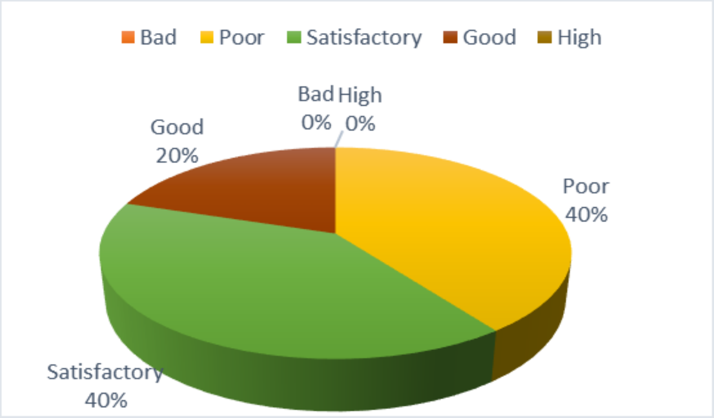
Table 12

The discriminating interpretation of essay test

No.	Interpretation	Item Number	Total Item	Percentage
1.	*Bad*	-	0 item	0%
2.	*Poor*	51 and 53	2 items	40%
3.	*Satisfactory*	52 and 55	2 items	40%
4.	*Good*	54	1 item	20%
5.	*High*	-	0 item	0%
Total			5 items	100%

Pie chart 4

Pie chart of the Discrimination Index in Essay test



Finding the result of this analysis, the researcher found that there were not items which were highly discriminate the students. The essay test was not being good at this categorized. It could be happen if the items were very easy to the students. So that, there were not items that could be discriminate between high and low test takers. The essay test here did not cover all material that should be measured. Al most all of items that the test maker made in essay test are about using grammatical accurately. It might be the reason why the discriminating did not have high discriminate level.

6. Distractor Analysis of the Test

It is important to evaluate the quality of each distractor in a test because a good distractor will attract more students from the lower group than the upper group and divert students from the correct answer if they do not know which is correct.

Anas sudjiono stated that, the distractor will be called work effectively if they have been chosen at least by 5% of all test participants.[96] The following table was the results of analysis of distractor analysis for number 1 and for the other items would use the same formula. The whole computation result of difficulty level can be seen in appendix.

Table 13

The Distractor Analysis of Multiple Choice

Multiple Choice Items							
No. Item	Group	A*	B	C	D	E	Total
1.	Upper	8	0	2	3	0	13
	Lower	1	1	9	1	1	13
Total		9	1	11	4	1	26
Interpretation		35%	4%	42%	15%	4%	

Note : the letter with a star marked is the key answer

There were only 8 students from upper group who answered ‘A’ and there were only 1 student from lower group who answered it. The answer key of item number 1 is ‘A’. Meanwhile, the others are the distractors. The distractor ‘B’ was chosen by only 1 student from lower group and none from upper group. The disctractor ‘C’ was chosen by only 2 students from upper group and 9 students from the lower group. While, The disctractor ‘D’ was

[96] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan*, (Jakarta: PT Raja Grafindo Persada, 2008), p.411

chosen by only 3 students from upper group and 1 students from the lower group and the disctractor ‘E’ was chosen by only 1 student from lower group and none from the upper group.

After computed all the distractors, the researcher concluded that the distractors ‘B’ and ‘E’ are not work effectively because they have been chosen by less than 5% of test takers. Then, the distractors ‘C’ and ‘D’ are work effectively because they have been chosen by more than 5 % of test takers. The table below is the result analysis of the distractor items.

Table 14
Analysis of the Distractor Items

No.	A	B	C	D	E	Total Effective Distractor	Total Ineffective Distractor	Total Distractor
1.	✸	X	✓	✓	X	2	2	4
2.	✓	✸	✓	✓	X	3	1	4
3.	✸	✓	X	✓	✓	3	1	4
4.	✸	X	✓	✓	✓	3	1	4
5.	✓	X	✓	✸	✓	3	1	4
6.	✓	✓	✓	✸	X	3	1	4
7.	✸	X	✓	X	X	1	3	4
8.	✸	✓	✓	✓	✓	3	1	4
9.	✸	✓	X	✓	✓	3	1	4
10.	✓	X	✸	X	✓	2	2	4
11.	✓	✓	✸	✓	X	3	1	4
12.	✓	X	✸	X	X	1	3	4
13.	✓	✓	✓	✸	X	3	1	4
14.	✸	✓	X	✓	X	2	2	4
15.	✸	✓	✓	✓	X	3	1	4
16.	✸	✓	X	✓	X	2	2	4
17.	✓	✸	X	✓	✓	3	1	4
18.	✓	✓	X	✸	✓	3	1	4
19.	X	✸	✓	✓	✓	3	1	4
20.	✸	X	✓	✓	✓	3	1	4
21.	✓	✸	X	X	X	1	3	4
22.	✸	✓	X	✓	X	2	2	4
23.	X	X	✸	✓	✓	2	2	4
24.	✸	X	X	X	X	1	3	4
25.	✓	✸	✓	X	✓	3	1	4
26.	✸	X	X	X	✓	1	3	4

27.	✓	✓	✓	✓	✸	4	0	4
28.	✸	✓	X	✓	X	2	2	4
29.	✸	✓	X	✓	✓	3	1	4
30.	X	✓	✓	✸	✓	3	1	4
31.	✸	X	✓	X	X	1	3	4
32.	X	✸	✓	X	X	1	3	4
33.	✓	✸	✓	X	✓	3	1	4
34.	✸	✓	✓	X	✓	3	1	4
35.	✸	X	✓	✓	X	2	2	4
36.	✓	✸	✓	✓	✓	4	0	4
37.	✸	✓	✓	✓	✓	4	0	4
38.	✓	X	✸	X	X	1	3	4
39.	X	✸	✓	✓	X	2	2	4
40.	✓	✓	X	✸	X	2	2	4
41.	✓	✸	✓	✓	✓	4	0	4
42.	✓	✓	✓	✸	X	3	1	4
43.	✸	✓	✓	X	✓	3	1	4
44.	X	✸	✓	✓	✓	3	1	4
45.	✸	✓	X	X	✓	2	2	4
46.	✓	✓	✸	✓	✓	4	0	4
47.	✸	✓	✓	✓	X	3	1	4
48.	✓	✓	✸	✓	X	3	1	4
49.	✓	X	✸	✓	X	2	2	4
50.	✓	✸	X	✓	✓	3	1	4
Total of the distractors						130 (65%)	70 (35%)	200

Notes :

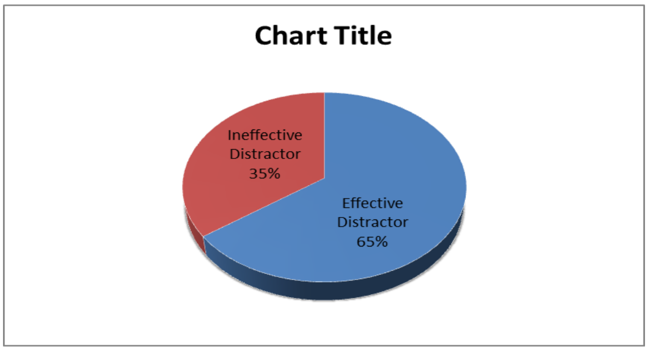
✸ = The key answer

✓ = Effective distractor

X = Ineffective distractor

Pie chart 5

Pie chart of the distractor analysis



Almost all of the distractor in the test had distractor which worked effectively. It could be seen in the final computation that there were only 35% distractor which did not work effectively. Doing further analaysis, the researcher found that it might be happen because the other option of correct answer (the distractor) in the test did not homogenized at all. Then, it also could be happen if the distractors were not logic and plausible. It was contrast with the basic principles of constructing the multiple choice test that the distractor should be homogeneneous, logic and plausible. For detail analysis could be seen in appendix.

C. Analysis

Analysis of the study consisted of determining whether the test is categorized of a good test. As the phenomenon found that there were many students in MA Uswatun Hasanah Semarang at the second grade of senior high school in the academic year 2013-2014 got low score in doing English summative test that made by MGMP LP Ma'arif NU of Semarang district. Here, the researcher tried to investigate what causes this problem by analyzing the test used.

The first analysis is content validity. The researcher analyzed content validity of the test and focused into 2 skills analysis; reading and writing skills, because the test that is used practically contained 2 skills; reading and writing. The test has content validity which is categorized adequate in term of content validity analysis. Based on the computation done by comparing between the contents of the test and the indicators recommended in the syllabus, the researcher found that the final result is 60,8%. It is regarded as adequate category because it is in the range from 56% to 75%. It means that the contents of the test have covered adequately to the indicators which are recommended in the syllabus.

Furthermore, the researcher analyzed the reliability, practicality, and the item analysis which consists of the difficulty level, the discriminating power and the distractor analysis. From this analysis, the researcher found that the reliability has fulfilled the requirements to be a good test. Based on the computation done, the test is reliable at both items test; multiple choice and essay test. It means that, the result of the test is the same even though the test administered to the same level students in different times. Practicality of the test has two criteria in simple and complete categories, but in complete categories the test did not included the scooring guidance aspect.

Meanwhile, the difficulty level of the test is in the rank of moderate level and the distractors work effectively proved by only 35% distractions that do not work effectively. It is the same as what Purwanto said, to be a good test, it should have moderate level and distractors which work effectively.

From six criteria explained to be a good test, just discriminating power which not fulfilled the criteria of good tests. The test did not have the high discriminating power and it will cause the difficulty to differ between upper and lower students. The cause of many students got low score may happen if the test is so difficult.

Meanwhile, the difficulty level of this test found that the test has moderate level. It means that, the test is not too difficult and not too easy and fulfilled category of good test. Then, the researcher concluded that the cause of many students got low score does not come from the test itself but it may come from almost all of the students do not master the material being measured yet.

D. Discussion

Test as an instrument has to have a good quality, because the quality of the test will influence the result of the test. If the test is good, the result will provide the right information to be used by the teacher in making accurate decision to the students' achievement. According to Robert L. Thorndike and Elizabeth Hagen, a well constructed test should have three main characteristics which involve validity, reliability, and practicality.[97]

[97] Thorndike, Robert L. and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Educatio*n, (New York: Jhon Willey&Sons, Inc.,), 2nd edition, p.160

A good test should also be good at its item analysis, that is some rather simple statistical ways of checking individual items. H. Douglas Brown stated that, “there are three main components of item analysis, they are: difficulty level, discriminating power and the effectiveness of the distractor.”[98]

The tests are called valid if they really measure what intended to measure. For example, when the teacher intends to know the capability of their students in mastering narrative text, the teacher just focuses on the material of narrative texts and they cannot input the other materials. Then, the test is called reliable if it can be used to test the students in the same level and other situations by having the same result. Meanwhile, the practicality of the test is the easyness to set the test up and to administere it.

Then, according to Purwanto, a good test item should have three criteria; moderate difficulty level, high discriminating power and distractor analysis which works effectively.[99]

1. Content Validity

The test had 60,8% of content validity. It was considered as adequate in its level, because it was range from 56% to 75%. It means that the contents of the test have covered adequately to the indicators which are recommended in the syllabus. The content of the test should cover all indicator in the syllabus, it is a must for constructing the test so that the test will be a good test. But, finding the result of analysis, then the researcher has a question in her mind, whether the MGMP arrange the test based on the basic principles of constructing the test. In the other hand, the teacher who joined in MGMP usually could not answer the questions, how the test was made, why the test only had adequate level in its content validity and the test did not cover all indicators in the syllabus. The case might be happened because of some reasons, the test maker just simplified when making the test. They did not make the test according to the basic principles of

[98] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004) p.58-60

[99] Purwanto, E*valuasi Hasil Belajar*, (Yogyakarta: Pustaka Pelajar, 2009), p. 99

constructing the test or the test maker might simplify when making the test because they had many busy things to do. Then, they only simplify when making the test and it made the test only had adequate level of content validity.

2. Reliability

The reliability of the test had only 0,989 in its multiple choice test and 0,806 in its essay test. It means that the test will have same result eventhough it is conducted in the other times with the same level students. The test did not had 100% of relibility it means that still there were some items which did not reliable. The researcher did not know whether the test had conducted before with the same result recently or may the test was the first time to conduct so, it still did not know the reliability itself. The test maker should think about it before they arranged the test. if the test had been conducted before by having imperfectly reliability they should revise the test items so that the test would have 100% of relibility. Anas Sudijono stated that, test maker should make the test in high reliability when arranging the test.[100]

3. Practicality

The Practicality of the test, the test maker did not include a scooring guidance to the teacher to score it. It made the teacher confused in scooring the test. Meanwhile, the test maker had given the key answer to the teacher but it did not for scooring guidance. Whether the test maker forgot to include a scooring guidance or might they really did not include a scooring guidance when making the test. As we know that MGMP is an organization which have many members. It was impossible they forgot together to include scooring guidance. Meanwhile, Anas Sudijono stated that, to be practical the test should have complete requirements, that it comes with instructions on how to do it, the answer keys and its scoring guidance.[101]

[100] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.97-99

[101] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.97

4. Difficulty Level

The difficulty level of the test had only 50% of moderate level in its multiple choice test and 60% of moderate level in its essay test. Knowing the fact that still there were the items test which did not in moderate level, the test maker should revised before the test was conducted. So that, it would not make students difficult in doing the test or vice versa. Then, items test will be really good. Why did not the test maker revise the items test, the researcher also has question in her mind. Whether the test maker might also simplify when making the test because they had many busy things to do. Then, they only simplify when making the test. They did not really correct the items and it made the test did not have high of moderate level. In the other hand, Purwanto stated that, a good test item should have moderate difficulty level.[102]

5. Discriminating Power

The discriminating power of the test had only 1% of high level in its multiple choice and none for essay test. The test maker should really revise the items test. Before conducting the test and arranging the test items, it a must for test maker to evaluate each items which donot suitable with the standard. So that, the test would really mesure students' achievement. The test could not differ low and high of test taker because it did not have high discriminating power which is the requirement to be good discriminating power. Doughlas stated that, discriminating power is the extent to which an item differentiates between high and low test-takers.[103]

6. Distractor Analysis

There were 35% distractor of the test did not function effectively. The cause of the test did not have good discriminating power because the construction of the test was not appropriate to the basic principles of constructing the test. The researcher found that it could be happen because the items of the test had options of the correct answer which did not homogenized. There were some options were

[102] Purwanto, E*valuasi Hasil Belajar,* (Yogyakarta: Pustaka Pelajar, 2009), p. 99

[103] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004) p.59

not plausible and were not logical. It would cause bad discrimiation items. Meanwhile, Damien Merchand stated that, the test items should be plausible, homogen and plausible.[104] The reason maybe the same as the content validity, it may be happened because the test maker only simplify when making the test and it made the test still had the distractor which did not work effectively.

E. Limitation of the Research

The research was limited by the document. The test was only analysis focusesed on its documents' file. The research did not explain whether how the test has been made before it was documented, how the reliable of the test before it was conducted, how the test items arranged and what the test really re evaluate by the test maker after it was conducted. Here, the research could not answer those questions because it was only focused on the documents' file.

[104] Merchan, Damien, *Basic Principles of Test Construction,* www.cte.cornell.edu, assessed at 14 December 2014.