The Quality of English Final Test at the Second Semester of Third Grade Students of SMAN 1 Pagaran in Academic Year 2016/2017

***Horhon Lumbantoruan**

****Sri Minda Murni**

****Isli Iriani Pane**

ABSTRACT

The objective of this study is to find out the quality of the English final test designed for the second semester of third grade students of SMAN 1 Pagaran in academic year 2016/2017. It describes whether or not the test items have good characteristic of test in terms of validity, reliability, difficulty level, and discriminating power. The test consists of 35 items multiple choice forms. The research design uses in this study was Descriptive Qualitative Research. To find out the discriminating power of the test, the writer chose the top 31% for the upper group and top 31% for the lower group. The result of this study shows that there are 18 (51%) acceptable items to meet the criteria of validity and 17 items (49%) is Invalid. The test is reliable since has 0.676 the level of validity. The test has unacceptable index of difficulty since has 15 items (43%) too difficult and are only 5 items (14%) easy items. Whereas for discriminating power index, the writer found there are 7 (20%) has negative result of the point have to be discard, 6 (17%) poor items, 8 (22%) satisfactory items, 13 (38%) good items, and 1 (3 %) excellent item. In conclusion, English final test designed for the second semester of third grade students of SMAN 1 Pagaran in academic year 2016/2017 does not meet the criteria of effective and acceptable test.

Keywords: Validity, Reliability, Level of Difficulty, Discrimination Power.

---

*Graduate Status
**Lecturer Status

INTRODUCTION

Background of the Study

Nowadays, evaluation takes an important role in education. Evaluation can measure the teacher' and students' progress in learning teaching process. The aim of evaluation itself is to evaluate students' achievement and students' progress in teaching and learning process. Testing is one of way to evaluate students' ability. Teaching learning process was successful or not can be seen from the result students' test.

Testing and Evaluation of language skills and competencies are very important components of language teaching. According to Chen Desheng and Ashita Varghese (2013:31), testing becomes an integral part of teaching because it provides significant information or inputs about the growth and achievement of learner difficulties styles of learning anxiety levels. It has an important role to measure the students' achievement during teaching and learning process. Thus, the test maker or test constructor should be able to construct a good test. Teachers who construct a good test will give a good contribution to student's education. On the other hand, teachers who have lack of skill in constructing a good test will give less contribution or might even make student's education become worst.

Dr. Foyewa, R.A (2015:32) stated that there are certain qualities expected of a good test. They include among others :

a. Validity : A good language test should measure what it supposed to measure. There are different types of validity. These are; face validity,

contents validity, predictive validity, concurrent validity and constructed validity.

b. Reliability: Reliability is the quality of being reliable: language test reliability is the consistency of a test in measuring what it is supposed to measure. A good language test is expected to be reliable.

c. Objectivity: this quality of a language test ensures that a test should have one and only one correct answer. Examples of this include the "multiple choice" and "true and false" tests.

d. Economy: this quality of test ensures that the cost of administering a test, the time involved in setting and marking it should be commensurate with the expected result obtained from it. A test that takes much time, energy and costs much to construct cannot be said to be economical.

A problem arises when most teachers underestimate an evaluation of the test item in the English final test they have made, whereas, this evaluation is in fact so important for the teachers in order to know the quality of the test they made; whether it is already valid and fulfill the characteristics of good test or not. Analyzing test items include analyzing the validity, reliability, level of difficulty, discrimination power.

Considering the explanation above, this study focuses to analyze the quality of English final test at third grade of students at SMAN 1 Pagaran concerning study about the validity, reliability, item analysis includes index of difficulty, and index of discrimination. The test is final school examination which constructed by the group of English teacher in SMA N 1 Pagaran. The form of the

test is multiple choice form. To analyze those aspects the writer could do either manually or using application of computer. Manually means to measure each of aspects the writer use formula stated or to make it easier and faster the writer can use the application of SPSS. In this study, the writer used SPSS version 17.

REVIEW OF LITERATURE

1. Testing, Measurement, and Evaluation

Testing, measurement and evaluation mean very different thing, and there mostly teachers has a wrong understanding to those words. Testing, evaluation, and measurement are three basic related concepts that we need to understand. The similarity among them is to assess the students' ability in mastering language. Test and measurement are parts of evaluation. The difference between test, evaluation and measurement can be found in the practice of assigning final marks to students at the end of a unit of work. For clear information about the differences among testing, evaluation, and measurement in the following writer will elaborate each of them.

2. Characteristic of Good Test

The teacher as test maker must be familiar with the characteristic of a good test in order to get complete information about the quality or the ability of the students in particular subject. All good tests made should definitely have good qualities, in this discussion there are some qualities to judge good test: validity, reliability, and item analysis includes index of difficulty, index of discrimination. In other words we could say, any kinds of test must be appropriate with terms of

four objectives (validity), dependable in the evidence it provides (reliability), the items not too easy and also not difficult (item difficulty), and the last the test should be able to differentiate students who has a high ability and low ability in mastering the subject (discrimination item). Without any one of them, a test would be poor in the quality. Whether the teachers are constructing their own test or selecting a standard instrument to use both in their class and school, they should certainly understand what these concepts mean and how to apply it.

a. Validity

Validity refers to measure accuracy it is intended to measure. According to (Hughes, 1989: 22) Validity in testing and assessment has traditionally been understood to mean discovering whether a test measures accurately what it is intended to measure. Validity consists of content validity, face validity, criterion-related validity (or predictive validity) and construct validity (Micheal &

b. Reliability

Reliability refers to the consistency of test result. Reliable here means that a test must rely and fit on several aspects in conducting the test itself. A test should be reliable toward students. J. Stanley Ahman said that many factors affect the reliability of measuring. Some are associated with the students themselves, whereas others factors arise from the instrument itself. Students factors ordinarily mentioned are state of health, fatigue, motivation, and emotional strain.

According to Heaton (1975:162), the factors affecting the reliability are:

1. The extent of the material selected for testing.

   Reliability is concerned with the size of the test; it is not too long and not too size.

2. The administration of the test.

   The students or test-takers must have same condition and time limit.

3. The instruction.

   The clarity of the instruction will affect the student's comprehension to answer the test.

4. Personal factors, such as motivation and illness.
5. Scoring the test

   It means that the objective test is more reliable than the subjective test.

c. Level of Difficulty

A good test is a test which is not too easy or vice versa too difficult to students. It should give optional answer that can be chosen by students and not to far by the key answer. Very easy items are to build in some affective feelings of "success" among lower ability students and to serve as warm up items, and very difficult items can provide a challenge to the highest-ability students (Brown, 2004:59). It makes students know and record the characteristics of teacher's test if the test given always comes to them too easy and difficult. Thus, the test should be standard and fulfill the characteristics of a good test. The number that shows the level difficulty of a test can be said as difficulty index (Arikunto, 2006:207). In this index there are minimum and maximum scores. The

lower index of a test, the more difficult the test is. And vice versa, the higher the test, the easier it is.

d. The Discrimination Item

It is the extent to which an item differentiates between high and low-ability test-takers. Discrimination is important because if the test-items can discriminate more, they will be more reliable (Hughes, 2005:226). It can be defined also as the ability of a test to separate master students and non-master students (Arikunto, 2 006:211). A master student is a student with higher scores of test, and a non-master student is a student with lower scores on the test given. The same as the term of difficulty level, discrimination has discrimination index. It is an indicator of how well an item discriminates between weak candidates and strong candidates (Hughes, 2005:226). This index is used to measure to the ability of a test in discriminating the upper and lower group of students. Upper students are students who answer with true answer, and lower group are students with false answer. In this index, it has negative point. Different from difficulty index, the negative index of discrimination power shows that the questions identify high group students as poor students and low group students as smart students. A good question is a question that can be answered by upper group and cannot be answered with true answer by lower group.

RESEARCH METHODOLOGY

In conducting the research, the writer used descriptive qualitative research design. This descriptive study was designed to obtain information concerning particular issues and then describe them. According to Donal Ary (2010:454) Qualitative descriptive study is the method of choice when straight descriptions of phenomena are desired. Descriptive research was presented a board of range of activities that have in common in purpose of describing situation and phenomena.

Method of Analyzing Data

To analyze the data, the writer measured the validity, reliability, item, difficulty, and the item discrimination.

Measuring the validity

To measure the validity, this study used formulas of *product moment*, the formula is written below:

$$r_{xy} = \frac{N.\sum XY - (\sum X)(\sum Y)}{\sqrt{\{N.\sum X^2 - (\sum X)^2\}\{N.\sum Y^2 - (\sum Y)^2\}}}$$

(Bachman, 2004:86 and Tuckman, 1978: 163)

2. Measuring the Reliability

To measure the reliability of the test, the writer used Kuder -Richardson 20 (Kr-20).

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{s^2 - \sum pq}{s^2}\right)$$

(Tuckman, 1978:163)

3. Measuring Index of Difficulty

Before analyzing the index of difficulty, the writer classified first the students' score made English teacher.

a. The steps to classify the students' score are:

1. Arrange the students' score of third grade class.
2. Then, arrange the students' score from high to low.
3. Then, classified 50% high score as upper group and 50% low score as lower group in each class.

b. To analyze the index of difficulty, the writer use Gronlund, (1993: 103) and Garrett (1981:363) formula as below:

$$P = \frac{R}{N} \times 100$$

Gronlund, (1993: 103) and Garrett (1981:363)

P = the percentage of examinees who answered items correctly.

R = the number of examinees who answered items correctly.

N = total number of examinees who tried the items.

A good test to be given the students is the test with the criterion of difficulty between 0,30 - 0,70. Meanwhile the index of difficulty which shows

0,00 - 0,30 and 0,70 - 1,00 was not good to be given to the students because the test either too difficult or too easy for them.

4. Measuring the Item Discrimination

The steps to analyze the item discrimination are:

a) Make table analysis to ease in analyzing index of discrimination.

b) Compute the index of discrimination using the formula:

$$ID = \frac{UR+LR}{N} \; x \; 100\ \%$$

(Gronlund, 1982:103)

D = Discriminating index

U = Number of students who answer the tem correctly in the upper group.

L= Number of students who answer the tem correctly in the lower group.

N = Number of students who took the test.

c) Match the result the item discrimination based on the criteria.

The criteria of item discrimination power could be seen in the below:

RESEARCH FINDING AND DISCUSSION

Findings

In this study the data were collected in qualitative research data. The qualitative findings were obtained from the analysis of the final test at the second semester for third grade of SMAN 1 Pagaran in academic year 2016/2017. The analysis deals with the analysis of validity, reliability, item of difficulty, and discrimination. It used the application of SPSS 17.

A. Analysis of Validity

4.1. Table of Validity

Item's Number	r Table	t Table	Validity level (r table> tTable)
1	0.465	0,325	Valid
2	0.352	0,325	Valid
3	0.070	0,325	Invalid
4	0.297	0,325	Invalid
5	0.254	0,325	Invalid
6	0.308	0,325	Invalid
7	0.304	0,325	Invalid
8	0.395	0,325	Invalid
9	0.349	0,325	Valid
10	0.375	0,325	Valid
11	0.493	0,325	Valid
12	0.018	0,325	Invalid
13	0.481	0,325	Valid
14	0.603	0,325	Valid
15	0.299	0,325	Valid
16	0.174	0,325	Invalid
17	0.621	0,325	Valid
18	0.072	0,325	Invalid
19	0.556	0,325	Valid
20	0.185	0,325	Invalid
21	0.471	0,325	Valid
22	0.189	0,325	Invalid
23	0.414	0,325	Valid
24	0.363	0,325	Valid
25	0.507	0,325	Valid
26	0.094	0,325	Invalid
27	0.048	0,325	Invalid
28	0.392	0,325	Valid
29	0.539	0,325	Valid
30	0.119	0,325	Invalid
31	0.385	0,325	Invalid
32	0.185	0,325	Invalid
33	0.021	0,325	Invalid
34	0.357	0,325	Valid
35	0.096	0,325	Invalid

Based on the analysis above, it can be concluded that the test items of the English final test for third grade in second semester of SMA N 1 Pagaran in academic year 2016/2017 are considered not achieved as a good test, because the measuring show that almost 50% of items INVALID.

B. Analysis of Reliability

Table 4.2 Analysis of Reliability

lCronbach's Alpha	N of Items
0.676	36

As the writer has stated in the previous chapter, the coefficient of reliability of test items is found by applying either the Kuder-Richardson 21 formula or by using SPSS application. From the computation, it is found that the coefficient of the test items is 0.676. As the criteria stated if the coefficient 0,60- 0,80 the test is reliable

3. Analysis of Difficulty Level

In analyzing the difficulty level of test, the students divided into three groups, they are upper group, middle group, and lower group. To divide the groups, first the researcher arranged the students' score from the high score into low score. Generally the score in the upper group is higher than KKM score applied, but in this study the highest score of the test only 66. So, in dividing those categories the writer took 31% the highest from the score, 38 % from the

middle score, and 31% from the lower score. The following table will show the results of data analysis of item difficulty level.

Table 4.3. Criteria of Index difficulty

Index of Difficulty	Criteria	Item Number	Total item
0.71-1.00	Easy	11,18,25,33,34	5
0.31-0.70	Moderate	2,3,4,5,13,14,15,16,17,22,23,28,29,31,35	15
0.00-0.30	Difficult	1,6,7,8,9,10,12,19,20,21,24,26,27,30,32.	15

From the analysis above, it can be conclude that the test items of the English Final Test at the Second Grade for Third Grade Students of SMA N 1 Pagaran in Academic Year 2016/2017 categorized as a bad test because the division of the three level of difficulty is not suitable. The test is need to be improved or revised in the next evaluation

4. Analysis of Discriminating Power

The index of discrimination is the ability to distinguish the students who achieve well or upper group and those who achieve poor or lower group. To analyze the index of discrimination, first the researcher arranged the students' score from the highest into the lowest.

The following table will show the result of the analysis of discrimination power of the final test.

Table 4.4 Criteria of Index Discrimination

Index of Discrimination	Criteria	Item Number	Total item
0.70≤ DP≤ 1.00	Excellent	4	1
0.40< DP≤ 0.70	Good	8,11,13,14,17,19,21,23,24,25,29,31,34	13
0.20< DP≤ 0.40	Satisfactory	1,2,5,7,10,15,22,28	8
0.00< DP≤ 0.20	Poor	6,20,26,27,30,33,35	6
-0	Wrong	3, 9,12,16,18,20,32	7

Based on the table above, the result of index of difficulty of discrimination power shows that there are 20 % (7 items) has a wrong index of discrimination. Those items are needed to be discarded. 17% (6 items) categorize as a poor items, those items should be revised. There are 8 items (22%) satisfactory in discrimination item. There are 38 % (13 items) good in discrimination, it needed a little revision and there is only 3 % (1 item) excellent in index of discrimination.

CONCLUSION AND SUGGESTIONS

Conclusion

Based on research findings, it is concluded that:

Based on the analysis of the 35 test items for third grade in SMA N 1 Pagaran in the academic year of 2016/2017, the following conclusions could be drawn:

a. In analysis of item validity, the English Final Test at the Second Semester of Third Grade Students of SMAN 1 Pagaran in Academic Year 2016/2017 has medium validity level because there are 18 items (51%) fulfill the requirements of validity and there are 17 items (49 %) that do not fulfill the requirements of validity.

b. By applying SPSS 17, the writer found that the coefficient of reliability of the whole test is 0.676. It means that the test is reliable, and we can use the test items as the instrument of evaluation again if we want to.

c. Considering from the level of difficulty, the writer found that the English final test made by the group of teacher in SMA N 1 Pagaran do not fulfill as the requirements as the good test or on the other words it categorize as a poor test. it found there are only 5 items (14%) out of 35 items are easy item, 15 Items (43%) moderate, and 15 items (43%) difficult. The number among: easy, moderate and difficult items are not balance.

d. In analysis of the item discrimination power, it was found that there are 7 items (20 %) has wrong index of discrimination, 6 items (17%) has poor index of discrimination, 8 items (22%) satisfactory index of discrimination, 13 items (38 %) of the items good discrimination items, and 1 item (3 %) excellent in discrimination item. We can conclude that the quality of the test based on the discrimination item is poor or it is not good.

Finally, the writer draws a conclusion that the test items in the English Final Test at the Second Semester of Third Grade Students of SMAN 1 Pagaran in Academic Year 2016/2017 is not good but it could still be used as an instrument of evaluation with some revisions.

Suggestions

Constructing good language test items (particularly objective test items) is not an easy task. Based on the conclusions above, the writer would like to offer the following suggestions.

a. For Teacher

Teacher as the test maker in the school he test constructors should know about the characteristics of good language test, especially the procedure of determining difficulty levels and discrimination power. To know whether the test is good or not, the test maker should try out the test first before testing it to the students. Items that can still be used should be revised and saved. The items that contain too many problems should be discarded.

b. For Students

The students should be careful in reading, analyzing, and answering the test. If they found the vague statements or questions , the students can ask the examiner to explain.

c. Other Researcher

The writer hopes the result on this item analysis could be used as an example in analyzing other test items, and encourages other researchers to do research on the same subject.

REFERENCES

Arikunto, suharsimi. (2009). *Dasar-Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara

Boopathiraj,C and K. Chellamani. (2103). *Analysis of Tests items on Difficulty Level and Discrimination Index in the Test for Research in Education.* International Journal of Social Science& Interdisciplinary Research. Vol.2.(2)

Brown, H Doughlas. (2004). *Language Assessment*: *Principles and Classroom Practices.* New York: Pearson Education

Desheng Chen and Varghese Ashitha . (2013). *Testing and Evaluati\on of Language Skills*. IOSR Journal of Research and Method in Education,Vol.1,2013,31-33. Bharathiar University Coimbatore.

Foyewa, R.A. 2015. *Testing and Evaluation in English Language Testing.* International Journal of English Teaching. Vol.3,No.6,32-40. European Center for Research and Training and Development UK.

Heaton, J.B. (1988).*Writing English Language Test.* New York. Longman

Nuryulia, Rini Ika. 2009. *Item Analysis of Achievement Test in Final Test for The Seventh Grade Students of SMP N 1 Moga Pemalang in the Academic Year of 2008/2009.*Thesis, Faculty of Languages and Arts UNNES, Semarang.

Tuckman, Bruce W. *Measuring Educational Outcomes, Fundamental of Testing*. USA: Harcourt Brace Jovanovich, Inc. 1975.