# AN ANALYSIS OF TEST ITEMS BASED ON THE CRITERIA OF GOOD TESTS

**Nur Sartika Putri, Ujang Suparman, Ramlan Ginting**

Nursartikaputri@gmail.com

## ABSTRACT

Masalah penelitian ini difokuskan pada kualitas butir soal yang digunakan dalam ujian semester. Tujuan penelitian ini adalah untuk mengetahui apakah kualitas butir soal bahasa Inggris sudah memenuhi kriteria soal yang baik atau tidak, terkait dengan validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan kualitas pilihan. Hasil menunjukan bahwa butir soal tersebut memiliki validitas yang baik sesuai dengan silabus, reliabiltas yang rendah (r=0.07), berdasarkan daya pembeda diketahui terdapat 10 soal yang jelek, 24 soal yang cukup, 12 soal yang baik, 3 soal yang negative, 1 soal yang sangat baik, dan berdasarkan tingkat kesukarannya diketahui terdapat 17 soal yang mudah, 16 soal yang sukar, dan 17 soal yang sedang.

The problem of the reseach was focused on the quality of test items used in semester exams. The objectives of the reseach were intended to determine the quality of of English semester test items whether or not fulfilled the following criteria of a good test: validity, reliability, discrimination power, level of difficulty, and the quality of options. The results of analysis proved that good validity because the material available in syllabus, low reliability (r=0.07), according to discrimination power were determined that 10 poor items, 24 satisfactory items, 12 good items, 3 negative items, 1 excellent item, and according to level of difficulty were determined that 17 easy items, 16 difficult items, and 17 average items.

**Keywords :** *criteria of a good test, iteman, test items analysis*

## INTRODUCTION

Testing refers to an effort to measure the result of student’s learning in teaching learning process. Consequently, the teachers should have an ability to arrange and to analyze a good test. Therefore, the accuracy and the carefulness of teachers may have a big impact on the increase of the quality of teaching particularly in giving the judgement of student’s ability. This information is very useful for both students in their learning and the teachers in their teaching. It can be a feedback for the teachers, who have responsibility to meet the instructional objectives, while for the students, it illustrates their performance.

Related to the importance of the evaluation, it is necessary to consider that the test should be well constructed. As a means of evaluation, a test is administered to get information about the student’s improvement and to measure the result of the teaching learning process. And semester test is a test activity which is held at the end of teaching learning process in one semester. That is why, the writer assumes that semester test is a kind of test which is intended as a feedback from the students and also as a result of teaching from the teachers in one semester. This information will be used to consider and to decide several rules not only for the student’s but also for the teachers in increasing the quality of teaching learning process. And the English test in Gedong Tataan is made by MGMP (*Musyawarah Guru Mata Pelajaran*). While MGMP itself consists of a team who has responsibility to design a test for each subject, it means that the semester test items are rarely analyzed by the teachers after they are tested.

To analyze the semester test items, there are some criteria of a good test according to some expert. A good test should have (1) Validity, (2) Reliability, (3) Level of difficulty, (4) Discrimination Power, and (5) The Quality of Options. This research was concerned with the whole with test items designed by MGMP. This includes test analysis and item analysis. Test analysis is administered to determine and describe such criteria as face validity, content validity, construct validity, and reliability. And the item analysis is used to determine about the level of difficulty, discrimination power, and the quality of options.

Shohamy (1985:3) supports that a test is a sample of knowledge and needs to be a good representation of it. It means that, what should be tested just a sample of behavior or knowledge, not the whole or behavior what the teachers has taught and the students have learned because it is also impossible to measure all of the students' abilities. The things that should be taken into account is the sample must be representative in the sense which is tested, it should reflect the knowledge that has been taught. The test that has been analyzed was achievement test and it was designed by MGMP. Achievement test tried to investigate the students' achievement based on the objective of a given material. Achievement test (Harrison as quoted by Hayatunnisa, 2003:8) tries to evaluate the test takers' language in relation to a given curriculum or material which the test-taker had gone through in a given course. It is intended to show the standard which the students have reached in relation to other students at the same stage.

A good test should fulfill certain the criteria. There are four criteria of a good test according to some expert; they are validity, reliability, level of difficulty, and discrimination power. Concerning about the criteria of a good test above, the writer was focused on the opinions.

Validity refers to the extent to which an instrument really measures the objective to be measured and suitable with the criteria (Hatch and Farhady, 1982:250). In other words, a test can be said to be valid to the extent that it measures what it is supposed to measure. If the test is not valid for the purpose for which its design, the scores do not mean what they are supposed to mean. Reliability refers to the consistency of measurement that is, to see how consistent test scores or other evaluation results are from one measurement to another (Grounlund, 2000:193). It means that a test is administered to the same condition on different occasion, the extent that it produces different result, it is not reliable. Discrimination power is an aspect of item analysis, discrimination power tells about which is the item discriminates between the upper group students and the lower group students. Shohamy (1985:81) states that discrimination index tells about the extent to which the item differentiates between high and low students on that test.

Difficulty level is one of kind of item analysis. Level of difficulty was concerned with how difficulty or easy the item for the students. Shohamy (1985:79) states that difficulty level relates to how easy or difficult the item is from the point of view of the students who took the test. It is important since test items which are too easy can tell us nothing about differences within the test population. If the item too easy, it means that most or all of the students obtained the correct answer. In contrast, if the item is difficult, it means that most or all of the students get it wrong. The quality of options is a distribution of testees in diciding alternatives on a multiple choice test. It is obtained by calculating the number of testees who choose the alternatives A, B, C, or D or those who do not choose any alternatives. From this way, the teachers would be able to identify whether distracters function well or bad.

## RESEARCH DESIGN

This research was intended to determine whether or not the first semester English test for the first year students of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year meets such criteria as validity, reliability, difficult level, discrimination power, and the quality of options. This research, the writer chose the first year students in the first semester of academic year 2012/2013 was observed. There were two classes of first years in the school, Computer Department and Automotive Department. Both of class used for research participants. To complete the data, the writer involved the English teachers and the experts as the second observe.

In analyzing the data, the writer used test analysis and item analysis. Test analysis serve as examination to evaluate the students. Test analysis was intended to analyze the whole test for determining the quality of the test, such as validity, and reliability. While Item analysis was a process which examines the students' response to individual test items in order to assess the quality of those items and of the test as a whole. And item analyses were utilized for investigating such criteria as difficulty level, discrimination power, and the quality of options. In analyzing the quality of option alternatives, the writer was used to ITEMAN and also as supporting data.

The procedure of this research carried out in some steps in test analysis and item analysis: In determining the content validity, the writer analyzed the test items by comparing the test items with Syllabus for the first semester of the first year of SMK. In calculating reliability of the test, the writer used KR 21.

$$Rt(KR21)=\frac{N}{N-1}\left(1-\frac{x(N-x)}{NS^2}\right)$$

N : the number of items in the test
x : the mean of the test scores
$S^2$ : the variance of the test scores
Rt : reliability

The correlation of coefficient was interpreted by using the following criteria:

0.90 – 1.00 : High
0.50 – 0.89 : Moderate
0.00 - 0.49 : Low
(Hatch and Farhady: 1982:247)

In calculating discrimination power (DP). The formula of discrimination power was as follows:

$$DP = \frac{U-L}{{}^{1}/{}_{2}\,T}$$

DP : Discrimination power
U : Upper group
L : Lower group
T : The total number of students
(Shohamy, 1985:81)

the criteria as follows:

DP : 0.00 – 0.20 is poor items
DP : 0.21 – 0.40 is satisfactory items
DP : 0.41 – 0.70 is good items
DP : 0.71 – 1.00 is excellent items
DP : Negative (Discarded, should be omitted) (Heaton, 1975:180)

While for calculating the level of difficulty (LD) of each item, The result of the addition divided by the two groups. The formula for computing the level of difficulty as follows:

$$LD = \frac{U+L}{T}$$

LD : The level of difficulty

U : Upper group who got the item correct
L : Lower group who got the item correct
T : The total number of students
(Shohamy, 1985:79)

The criteria:

LD : 0.00 – 0.30 is difficult
LD : 0.31 – 0.70 is average
LD : 0.71 – 1.00 is easy
(Shohamy, 1985:79)

## RESULT AND DISCUSSION

### Results

a. Validity

The validity of English semester test items administered at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year was catagorized as good because the fifty objective test items represented on the material available in English Curriculum 2006. It means that most of the items in the achievement test were in line with the theory of language.

b. Reliability

The reliability of English semester test items by using formuation Kuder Richardson 21 or KR 21. The coefficient of the reliability was 0.07. Based on the criteria, it means that reliability of English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year has low reliability.

c. Discrimination power

Discrimination power of the test was calculated by using formula. The result shows that there were 10 items considered poor, 24 items were satisfactory, 12 items were considered good, 3 items were considered negative, and one items were excellent.

d. Level of difficulty

The writer added the number in upper group who got the items correct and lower group who got the items correct then divided by the total number of students. There were 17 items which were easy, 16 items which were difficult, and 17 items which were average. Based on the result, it is known that 3 items should be discarded because they have negative discrimination powers. There were 25 items which should be revised since item number 16 has poor item and level of difficulty was easy, and item number 3 also has good item but level of difficulty was difficult. Furthermore, item number 43 has also poor item but level of difficulty was average, 8 items were poor items and its level of difficulty were difficult, 5 items were satisfactory items but its level of difficulty were difficult, and 9 items were satisfactory items but its level of difficulty were easy. However, there were 22 items which were acceptable to be used. 10 items were satisfactory items and the level of difficulty was average. Then, 4 items were good items and the level of difficulty was average, 7 items were good items and the level of difficulty was easy, and one item was excellent items and the level of difficulty was average.

e. The Quality of Options (Prop Endorsing)

The quality of options the English semester test items of the test was analyzed by using ITEMAN. From the result of the quality of options analysis by ITEMAN, it concluded that there were 15 key answers needs revision because the other alternative is a better answer than the key answers. According to Ngadimun

(2004:10) it means that the discrimination power was getting by the students in upper group cannot answers the test incorrect, but the students in lower group can answer the test correctly (perhaps just right). In addition, it was found that Alpha (reliability based on ITEMAN) was 0.430, standard deviation is 3.645, variance is 13.287 and average mean of each item is 0.176 and mean biseral (average mean of the whole test items is 0.243, furthermore, it was found that there were 29 alternatives have rejected, 60 alternatives have accepted, and 111 alternatives have revised from 200 option alternatives.

## Discussion

In this research, before the writer gave the semester test to the students, the teacher gave a suggestion in order to the writer used both of class as research participants. At the meeting, the writer gave the English semester test and the writer firstly explained about the instruction of semester test. It was intended to make them easier to understand about the instruction of the test. The writer gave the students 90 minutes to answer the English semester test items.

And based on the result of students' questionnaire of face validity, it was obtained that the calculation was 69,3 %. It indicates that face validity of semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year was catagorized as good. It means that face validity fulfilled the criteria of a good test.

Based on the result of content validity, it was obtained that the content validity of semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year in the English subject have had a good content validity. According to English curriculum 2006 for vocational high school, there were 5 aspect of English being taught, the English semester of Listening were 30%, reading test item were 30%, vocabulary were 4%, language function were 28%, and writing were 8%. It means that the entire English semester test items was used in the test has been studied in for the the first semester of the first year of Vocational High School. Shortly, comparison of the number in the test items particularly in the English semester test for each material was already representative. For instance, According to English curriculum 2006, the English semester of reading test was focused on how to find out identify and respond an invitation. And the result of the objective found that 5 items or 10% about invitation.

Based on the result of construct validity of the English semester test items, the result shows that construct validity of the semester test items was 75%. Based on the criteria of construct validity, the semester test items for the first year students of SMK Negeri 1 Gedong Tataan was good. It means that most of the items in the achievement test were in line with the theory of language.

Based on the result of reliability of the English semester test items, the writer was utilized to manual data in calculating reliability, the result shows that the highest score was 33 and the lowest score was 14 from 50 English test items. Reliability of the test was 0.07. It means that the test was catagorized as low, while the criterion for low reliability is in range 0.00-0.49.

**Table 1 Table of the result of Discrimintaion Power**

| Criteria | Classification | Test Items |
|---|---|---|
| 0.00 – 0.20 | Poor items | 10 items |
| 0.21 – 0.40 | Satisfactory items | 24 items |
| 0.41 – 0.70 | Good items | 12 items |

| | | |
|---|---|---|
| 0.71 – 1.00 | Excellent items | 1 item |
| Negative | Discarded, should be omitted | 3 items |

The results shows that there were 10 items considered poor items, 24 satisfactory items, 12 good items, 3 negative items, and one excellent item.

**Table 2 The result of level of difficulty**

| Criteria | Classification | Test Items |
|---|---|---|
| 0.00 – 0.30 | Difficult | 16 items |
| 0.31 – 0.70 | Average | 17 items |
| 0.71 – 1.00 | Easy | 17 items |

The result of level of difficult, there were 17 easy items, 17 average items, and 16 difficult items.

Based on the result of the English semester test were related to criteria of the level of difficulty and discrimination power, it can be inferred that 22 items were administered, 25 items were revised and 3 items should be dropped because it have negative discrimination power. It can be seen from the level of difficulty and discrimination power above.

Based on the result of the quality of options (prop endorsing) by using ITEMAN, it shows that there were 15 key answers needs revision because the other alternatives have good chance better than the key answers have been fixed. There were 29 alternatives have rejected, 60 alternatives have accepted, and 111 alternatives have revised from 200 option alternatives. As example, here there was some item which has problems in listening section, reading section, error recognition, and reading comprehension. In listening section, there were 6 key answers needs revision. Perhaps, one of the matters is the students can not listen what the speaker said clearly. As example, number 11 was about short conversation. In this part, the students heard short conversation, and the students heard the conversation twice.

*What is wrong with the man?*
*a. He is busy on Tuesday*
*b. He doesn't like a doctor*
*c. He doesn't feel well*
*d. He is overweig*

The conversation:

A: "I have a bad fever, I need to see a doctor.'

B: "The doctor can see you until Tuesday."

A: "Tuesday? I can't when until then."

From the test above, the result of the quality of options by using ITEMAN as following:

| Item statistics | | | | Alternative statistics | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Alt | Prop. Endorsing | | | Point Biser | Key |
| No. | Prop. Correct | Disc. Index | Point Biser | | Total | Low | High | | |
| 11 | 0.04 | -0.05 | -0.39 | A | 0.39 | 0.42 | 0.38 | 0.03 | |
| | | | | B | 0.57 | 0.53 | 0.63 | 0.13 | ? |
| | | | | C | 0.04 | 0.05 | 0.00 | -0.39 | * |
| | | | | D | 0.00 | 0.00 | 0.00 | 0.00 | |

Item test numbers 11, key answer C, interpretation:

There was the information from his item test *'Check The Key, C was specified, B works better'*, it means that the alternative answer B is a better answer than C. Perhaps the key answer (C) needs check anymore, it proved that the *point biser* (discrimination power) in this item shows that -0.39 (*point biser* is very low or very poor, because D ≤ 0.199). The *point biser* can be interpreted that the smart students cannot answer the item test correctly, but the low students can answer the item test correctly or it just coincidentally. Based on the ITEMAN, The *Prop correct* of this item was 0.04, it means that the item test was 'difficult' (p<0.25).

And in the following paragraph, the classification of analysis of alternative answers the item test number 11 based on ITEMAN.

a. Alternative answer A, *prop. correct* was 0.39, it means that the test was classified into ‘average’ and therefore, it was good item. *Point biser* in this alternative answer was 0.03, it means that the test was classified into very low and it needs total revising.

b. Alternative answer B, *prop. correct* was 0.57, it means that the test was classified into ‘average’ and therefore, it was good item. *Point biser* in this alternative answer was 0.13, it means that it needs revising. But the smart students choose this alternative more, perhaps because this item was listening test, so the students was just answer coincidentally.

c. Alternative answer and key answer C, *prop. correct* was 0.04, it means that the test was classified into ‘very difficult’ and therefore, it needs total revising. *Point biser* in this alternative answer was -0.39, it means that it needs total revising or dropping. Because the low students choose this alternative more.

d. Alternative answer D did not have function as distractor, because the whole of the students did not choose the alternative answer. If it happens, so the alternative answer D needs total revising or dropping.

From the results of the research by using ITEMAN, it make easier to analyze the data than manual data. Based on manual data and ITEMAN, it can be concluded that the quality of English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 needs revising because it have low reliability although  face validity, content validity, and construct validity were catagorized as good. In this case, it would be better for the committee of

Dekdikbud who made the test to analyze it after administering to the students. Then, the teacher should have the capability to evaluate anymore the semester test items in order to find the student's weakness. Thus, the students get more concerned about answering the question clearly. Actually, in calculating data by using manual data and ITEMAN were effective but by using ITEMAN in calculating data, it was more easy, fast, and efficient. It was recommended for the teachers to try using ITEMAN software or the other software such as SPSS (Statistical Program for Social Science), Anates, and Microsoft Excel to analyze the result of the item test their students.

## CONCLUSIONS

Based on the results of the data analysis, the test has been analized was achievement test, in case of semester tests, which was designed by MGMP Gedong Tataan, Pesawaran. Achievement test investigated the students' achievement based on the objective of a given material and the writer draws the results of analysis proved that it has good validity because the material available in syllabus, low reliability (r=0.07), according to discrimination power were determined that 10 poor items, 24 satisfactory items, 12 good items, 3 negative items, 1 excellent item, and according to level of difficulty were determined that 17 easy items, 16 difficult items, and 17 average items.

## BIBLIOGRAPHY

Hatch, E, and H. 1982. *Research Design and Statistic for Applied Linguistic*. London: New Burry House, inc.

Hayatunnisa, 2003. *An Analysis of the First Semester English Test for the Second Year Students of SMU AL-KAUTSAR Bandar Lampung*. Unpulished Script. Bandar Lampung: FKIP University of Lampung.

Henning, G. 2000. *A Guide to Language Testing*. Mass : New Bury House Publishers.

Shohamy, E. 1985. *A Practical Handbook in Language Testing For the Second Language Teacher*. Tel Aviv: Tel Aviv University.

Ngadimun. 2004. *Analisis Butir Soal dengan Komputer dan Menafsirkannya*. Makalah disampaikan pada sosialisasi KBK bagi guru SMP Kabupaten Tanggamus di Pulau Panggung, tanggal 22-24 Juli 2004. Bandar Lampung:HE