KAIROS ELT JOURNAL, Vol. 4 No. 1 April 2020
Copyright©2019, ISSN: 2580-4278

TEST ITEM ANALYSIS OF READING COMPREHENSION EXAMINATION FACULTY OF TEACHERS AND TRAINING EDUCATION

Viator Lumban Raja
Catholic University of Saint Thomas
Email :viator_lumbanraja@ust.ac.id

ABSTRACT

It is not uncommon to put a blame on the students when they fail in the semester examination. The examiner or the one who constructs the test is rarely blamed or questioned why such a thing can happen. There is never a question whether the test is valid or reliable. In other words, the test itself is never evaluated in order to know if it meets the level of difficulty and power of discrimination. Madsen (1983: 180) says that item analysis tells us three things: (1) how difficult each item is, (2)whether or not the question discriminated or tells the difference between high and low students, (3) which distracters are working as they should. This reading comprehension examination consists of 44 items, 35 items of reading comprehension and 9 items of vocabulary. The number of test takers are 18 students. The result of the analysis shows that only 5 students (27.7%) can do the test within average, meaning they can answer the test 50% correct of the total test items. This belongs to moderate category, not high nor excellent. Of the 44 test items, 33(75%) are bad items in that they do not fulfill one or both of the requirements concerning the level of difficulty and power of discrimination. And only 11 items (25%) meet the requirements of level of difficulty and power of discrimination. Regarding the distracters, there are 20 items (45.45%) whose distracters are not chosen either one or two. There are two items (4.54%), 25 and 34, the correct answer of which is not chosen by the test takers, including the high and low group. In short, these 20 items needs revising in term of distracters. Revision is made to those items whose distracters are not chosen and those which do not fulfill the requirements of level of difficulty and power of discrimination. Distracters which look too easy are changed, and those which are not totally chosen are revised.

Keywords: *item analysis, level of difficulty, power of discrimination, effective distracters*

A. Introduction

1. The Background of the Study

Academic Activities certainly relate with examination in order to evaluate the learning-outcome of the students. In universities, examinations are administered at least twice in a semester, they are mid-term examination and final-term examination. Therefore, in a year there must be four times examination because a year has two semesters, they are odd and even semester. One semester can last twelve to fourteen weeks. If the meeting does not still fulfill the requirement of the meeting, examination is not allowed to be administered. Thus, a lecturer should fulfill the number of the meetings in order to administer the examination.

Failure certainly disappoints the test-takers. Very seldom does it occur that teacher is the one to blame if more than a half of the number of the students fail in the test. The teacher always avoids being got wrong by saying that the students do not prepare themselves well for this test. The question is, can it be justified that the students are the ones to blame if most of them fail in a test? Perhaps, the test itself needs analyzing whether it suits the instructional objectives or not. In other words, the test material should match the teaching material which has been taught to the students. Besides, the test items should also match the stated objectives, otherwise they do not measure what they are intended to measure. In relation to his, they learning outcomes need to be measured in order to know whether or not they accomplish the instructional objectives.

However, rarely do we evaluate the examination itself. Does it meet the requirements of a good test? Does it measure what it is intended to measure? Does the result of the test indicate consistency? Does it match the syllabus? Does it examine what has been taught? Such questions never arise among the lecturers although some students may complain about it. The students' complaint that the test does not match with what they have been taught is rarely heard. Possibly, the test item is not clear for students to understand so that they give the answer far away from what the lecturer has expected. In this case, the students are not the only ones to be blamed, but the lecturer as well because the test items is not clearly stated or it sounds ambiguous.

Examinations are surely intended to evaluate the students' learning outcome after carrying out a program in a certain period of time. The learning outcome, however, can be unsatisfactory if the instrument used to measure does not meet the criteria of a good test. Therefore, a bad result of an examination cannot be totally burdened to the students only, but the instrument itself should be evaluated whether or not it meets the precise requirements as a good test. A test is said to be good if it is valid, reliable, economical and interpretable (Tuckman, 1975).

The way how to measure the learning outcome is to do an evaluation. According to Grounlund (1985: 5), evaluation is the systematic process of collecting, analyzing and interpreting information to determine the extent to which students are achieving those instructional objectives. In regard to evaluation, tests are usually designed to measure the intended learning outcomes, especially those concerning the cognitive domain. Though the affective and psychomotoric domains are also mentioned in evaluating the students, the cognitive domain is more emphasized since it can be clearly seen from the result of the test. Thus, ideally, the instructional objectives will state the desired changes and the test as an evaluation instrument will measure the extent to which those changes have taken place. In fact, an evaluation may consist of measurement and non-measurement. The former refers to test which provides quantitative data or numerical data, while the latter refers to qualitative data such as attitude, participation during the learning-teaching process, percentage of attendance, etc. So far, our evaluation in this university belongs to the measurement, that is the test which is administered twice in a semester.

Of the four language skills, reading comprehension is in the third rank which also needs a full attention in carrying it out. As a matter of fact, reading comprehension can be seen as interactive process between readers and the writer through a text which leads to comprehend the content of the text (Alyousef, 2005). Reading comprehension test may cover a wide variety of topics. In general, test

items of reading comprehension are arranged in multiple choice form in order to be objective in scoring. But to construct multiple choice form test for reading is not as easy as that of essay.

Therefore, this study is intended to examine the test items of reading comprehension including vocabulary administered to the students of the academic year 2014/2015. What is going to be carried out here is to examine and analyze each test item which the students answered. This study is focused on judging whether or not each test item fulfils the requirements of level difficulty and power of discrimination. Besides, each distracter is also analyzed whether or not it is effective. Thus, this study will not give value judgment of the students' ability on reading comprehension, but on the test items and their distracters.

2. The Problems of the Study

In regard to the background above, the problems of this study can be formulated as follows: (1) Does each item test meet the requirements of item difficulty? (2) Does each item meet the requirements of power discrimination? (3) Are the distracters effective? The objectives of the study is to answer the stated problems above.

To calculate the co-efficient of these two problems, the formula proposed by Tuckman (1975:272) is accurately applied.

B. Review of Literature

The learning-teaching process which occurs in a certain period of time needs evaluating in order to know whether the instructional objectives are accomplished. If, for example, the percentage of the students who achieved the passing grade is only forty percent, that means instructional objectives are not accomplished because it is below the target. To say, for instance, there must be at least seventy percent of the student numbers who reach the passing grade. It sticks out of a mile that there are a plenty of factors which might make the instructional objectives not achieved such as intelligence factor, students' motivation, and the test itself. The two aforementioned factors are hard to detect or investigate, and what seems possible to investigate is the test. Some questions can be raised about it: is it a good test? Is it valid and reliable? Is it economical? Do the items have item difficulty and discrimination power? A series of questions can be raised in order to examine the test itself. This is necessary to do in order to know whether or not a test meets the requirements as a test. One cannot assume that his test is good if he does not get a feedback from others who evaluate his test. As a teacher, one should evaluate his test from time to time so that he can have a valid and reliable test. If a test has fulfilled the validity and reliability, there is no doubt that the test has already met the requirements as a good test.

1. Evaluation and Measurement

Very often do we get confused with the term 'evaluation and 'measurement' in the learning-teaching process. Which do we administer during the semester examination? According to Tuckman (1975) evaluation is a process where in the parts, process, or outcomes of a program are examined to see whether they are satisfactory, particularly with reference to the program's stated objectives. Thus, evaluation covers a lot of factors including the testing which is only one type of

evaluation. On the other hand, Gronlund (1985:5) indicates that evaluation is the systematic process of collecting , analyzing and interpreting information to determine the extent to which students are achieving the instructional objectives. Information may be obtained from informal classroom observation.

Measurement belongs to test instrument which is a form of evaluation. Therefore, evaluation comprises two kinds, measurement and non-measurement like rating scale while measurement is testing. Tuckman (1975: 3) states that the importance of tests as an integrated part of educational process is not only for monitoring students progress, but also for diagnosing strength and weakness of our students, or it is a tool for finding out what our students have learned (Copperud,1979). Therefore, tests should result in terminal-required objectives. Thus, within tests, a teacher is supposed to be able to identify what aspects of the instructional teaching materials which the students have not yet mastered, and which ones they have. In addition to that, the tests should be able to measure what we intend to measure and their results should be consistent.

A good test requires several conditions which should be fulfilled as well as possible. However, if it is standardized tests, they must have already had validity and reliability such as TOEFL, IELTS, TEFL, ALIGU, etc.

As it has been mentioned before the test which is under discussion in this study is a standardized test the validity and reliability of which is already fulfilled. On this occasion, the study focuses on the item difficulty and discrimination power as well as its distracters whether they are effective or not.

2. Item Analysis

Item analysis refers to each item of the tests which is thoroughly analyzed in order to know if the distracters are effective or not. It is an analysis of the relationship between item scores and the total test scores which often reveals those items that are inconsistent with the total test or parts of it. In fact, item analysis is the procedure by which individual item performance by a group of test takers is compared to their performance on the total test (Tuckman, 1975: 271). It will give a picture about the level of item difficulty, discrimination power, and effectiveness of distracters. To get the effectiveness of each item test, it can be determined by analyzing the students' response. This item analysis is usually designed to answer questions as the following: (1) Does the item function as intended? (2) are the test items arranged within appropriate difficulty? (3) are the test items free of irrelevant clues and other defects? Since the test under discussion is objective test with multiple choice answer, the three questions above refer to the choice provided which is based on the stem. In short, each item should function as it is intended, be arranged with appropriate difficulty in accordance with students' capability and teaching materials they have studied, and all test items with their choices should be free of irrelevant clues and defects. This means there must be a relationship between the stem and the choices provided which must be free from irrelevant and grammatical mistakes.

Answer to such questions are of obvious value in selecting or revising items for future use (Gronlund, 1981: 244). So far he gives the benefits to the importance of doing item analysis tests. The benefits are not limited to the improvement of individual test items, but there are a number of fringe benefits of special value to classroom teachers as well. The most important of this item analysis is that this data

provides a basis for (1) efficient class discussion of the test results, (2) remedial work for those who fail the test, (3) the general improvement of classroom instruction. In addition, Madsen (1983: 180) says that item analysis tells us three things: (1) how difficult each item is, (2) whether or not the question discriminates or tells the difference between high and low students, and (3) which distracters are working as they should. However, according to Saleemi (1988) item analysis is to look at the result of a test in terms of item difficulty and item discrimination. As a matter of fact, all these ideas are interrelated because if one test item is not correctly answered by all the test takers, then it has no item difficulty and at once it does not discriminate the high and low students. This indicates that one test item with high difficulty must be correctly answered by more students of the high group, and few of the low group. If for example, one test item with high difficulty is correctly answered by four students of the high group, then one student of the low group, then the test item must meet the level of difficultly and power of discrimination. But if no one answers that test item, it should be discarded or totally revised. The number of the high group and low group should be equal.

a) Item Difficulty

Item difficulty or level of difficulty refers to an item which is correctly answered by the test takers. If a test item can be correctly answered by all the test takers, that item has no item difficulty because all test takers can answer it correctly. On the other hand, if a test item cannot be correctly answered by the test takers, that item has no level of difficulty either. The former item is so easy that every test taker can answer it, whilst the latter item is so difficulty that no test taker can answer it. Therefore, a test item can be said to have level difficulty if there is a proportionally correct answer given by the high and low group of the test takers.

To determine high and low group of the test takers, first of all we separate 25 percent of the total scores after being ranked as the highest to the lowest score. Say, for example there are 20 test takers for reading comprehension. Then, their scores are ranked from the highest to lowest score. Next, we separate 25 percent, and that will be 5 students, rounded to 6 because there must be even number for the high and low group. As a result, there are 3 persons representing the high group and 3 persons representing the low group. The formula applied to calculate the coefficient of item difficulty is as follows: (Tuckman, 19975: 272).

$$\frac{\text{Correct answers by high group + those of low group}}{\text{Total number of high and low group}}$$

We can find the item difficulty of item no.1, for example, for the test of reading comprehension with high and low group 3 person respectively. The high group answers the item correctly, while only one student of the low group answered it. That item is already proportionally answered, and it has the item of difficulty of 0.67. Item difficulty index ranges from 0.30 – 0.90.

b) Item Discrimination

Item discrimination or discrimination power is in positive direction if more test takers in the high group than low group get the item right. It indicates that the item is discriminating in the same direction as the total test score. Since we assume that the total test score reflects the achievement of desired objectives, we would like

all of our test items showing positive direction. The discrimination power of an achievement test item refers to the degree to which it discriminates among the test takers with high and low achievement. If one test item, for example, is correctly answered by more students of high group than low group, it belongs to good item test. On the other hand, if it is correctly answered by both groups, then that item has now power discrimination. Discrimination power is in negative direction if more students in the low group than the high group get the item right. Therefore, an item with no power discrimination is one in which an equal number of test takers in both the high and the low group get the item right. On the contrary, an item with maximum positive discrimination power is one in which all test takers in the high group get the item right, and all the test takers in the low group get the item wrong.

Tuckman (1975) further said that item discrimination power has minimal index of 0.40. This means that if it is lower than 0.40, that items is poorly made, and it should be revised or discarded because it cannot distinguish between high and low group of the students. Likewise, item difficulty with index of lower than 0.30, it should be revised or discarded since it has no level of difficulty both for the high group of students and low group of students. The following is the formula how to calculate the coefficient of discrimination power:

$$\frac{\text{Correct answer of high group}}{\text{Number of the high group}} - \frac{\text{those of low group}}{\text{number of the low group}}$$

c) Effectiveness of Distracters

How well each distracter is operating can be determined by inspection of each item, and there is no need to calculate an index of effectiveness although the formula for discrimination power can be used for this purpose., In fact, it can be known from the answer of both high and low group. Again, if an item can be totally answered by both high and low group, that means that distracter is poorly constructed because there is only one choice picked up by the high and low group. The distracters should be picked up both high and low group.

In general, a good distracter, not the correct one, attracts more test takers of the low group than the high group. Thus, it should discriminate between the high and the low group in a manner opposite to that of the correct alternative. An examination of the following item-analysis data will illustrate the ease with which the effectiveness of distracters can be determined by inspection.

Alternative/distracters	A*	B	C	D
High Group = 10	5	4	0	1
Low Group = 10	3	2	0	5

*correct alternative

Based on the data alternative B is a poor distracter because it attracts more students from the high group than the low group. This is most likely due to some ambiguity in the statement of the item (Grounlund, 1985: 250). Alternative C is completely ineffective as a distracter because it attracts no one either from the high group or the low group. Alternative A as the correct alternative functions as intended because it attracts more students from the high group than the low group. Likewise, alternative D functions as intended because it attracts more students from

the low group than the high group. To sump up, this item just needs a little revision of alternative C which is not selected by both groups.

C. Research Method

This research belongs to descriptive research which concerns with conditions or relationship that exits. It describes and interprets what is (Ary, et.al, 1977). Precisely, this is a quantitative descriptive research because it describes numbers or figures and tries to interpret why such a phenomenon happens.

In fact, descriptive research is a bit similar to qualitative research as it concerns with existing phenomena, using data which may have been collected or taken from available sources such as the student records, the students' test result, students' academic achievement or students' social condition. The conclusion drawn from this descriptive research is not as strong as that from experimental one since there is no intervention of both independent and dependent variables attached within it. In descriptive research there is no generalization because it does not concern with population and sample. Although it is said quantitative descriptive, the "quantitative" here refers to the number of the test takers who choose alternative A, B, C or D of the reading comprehension examination. This number is required in order to determine the level of difficulty and power of discrimination of each test item. In addition, that number is also required in order to decide whether or not each distracter effective.

The data of this research are taken from the students' test result of reading comprehension examination which has been available in the administration office. The reading comprehension examination was administered to the students of English Department Faculty of Letters in 2015 and the result was well kept in the office. This reading comprehension examination consists of 35 items, twelve of which are cloze passage, and 9 items of vocabulary. There are 44 test items altogether.

The number of the test takers are 18 students, and that makes 18 answer sheets to be identified, classified and interpreted. Then, each item of the test is analyzed to find out the level of difficulty and power of discrimination. To find out the effectiveness of each distracter, each test item should be recorded how many test takers choose each distracter of it. A distracter is said to be effective if it is chosen by the test takers. If it happens that one or two distracters of a test item are not chosen or left blank, that means the distracter is not effective.

First of all, the score of the test result is ranked from the highest to the lowest in order to decide the high group and the low group. These two groups will decide the level of difficulty and power of discrimination through a certain formula. A coefficient of 0.31 – 0.92 is categorized good for level of difficulty, and 0.40 – 0.70 is average, 0.80 – 1.00 is categorized good for power of discrimination.

As for the effectiveness of distracters it should be identified which distracter of each test item is not chosen by the test takers. Each test item has four alternatives: A, B, C and D. If one distracter of a test item is not chosen but it still meets the requirements of level of difficulty and power of discrimination, then it does not need revising or changing because the number of the high group is bigger than the low group in choosing the correct answer. On the other hand, if one distracter of a test item is not chosen and it loses the power of discrimination because the number

of the low group is bigger than the high group in choosing the correct answer, then the distracter or the lead needs revising or changing.

D. Finding and Discussion

1. Item Analysis

a) Level of difficulty

A test item is said to have a level of difficulty if more of the high group answer it correctly than those of the low group. On the contrary, if a test item is correctly answered by more of the low group than the those of high group, then that item has bad power of discrimination. If a test item is correctly answered by the same number of both low and high group, then that item has no power of discrimination. It means it cannot distinguish betweens the low group and the high group.

Below is the result of test item analysis referring with level of difficulty and power of discrimination.

Item	Correct Answer	High Group	Low Group	Level of Difficulty	Power of Discrimination	Recommendation
1	D	2	3	0.83	-0.34	Revised
2	C	3	1	0.66	0.67	Accepted
3	D	3	2	0.83	0.34	Accepted
4	B	3	2	0.83	0.34	Accepted
5	D	1	3	0.66	-0.66	Discarded
6	B	2	3	0.83	-0.34	Discarded
7	B	3	3	1	0	Discarded
8	B	2	2	0.66	0	Discarded
9	D	3	2	0.83	0.34	Accepted
10	B	2	0	0.33	0.66	Accepted
11	C	3	2	0.83	0.34	Accepted
12	B	3	3	1	0	Discarded
13	B	2	2	0.66	0	Discarded
14	C	2	2	0.66	0	Discarded
15	A	3	3	1	0	Discarded
16	C	1	3	0.66	-0.66	Discarded
17	A	2	3	0.83	-0.34	Discarded
18	D	1	1	0.33	0	Discarded
19	C	1	0	0.16	0.33	Revised
20	B	1	0	0.16	0.33	Revised
21	C	2	0	0.33	0.66	Accepted
22	B	1	0	0.16	0.33	Revised
23	B	2	2	0.66	0	Discarded
24	B	1	1	0.33	0	Discarded
25	D	0	0	0	0	Discarded
26	C	2	1	0.55	0.33	Accepted
27	C	0	1	0.16	0.33	discarded
28	B	3	2	0.83	0.34	accepted
29	A	1	2	0.5	-033	Discarded

30	D	1	1	0.33	0	Discarded
31	A	1	2	0.5	-0.33	Discarded
32	B	1	0	0.16	0.33	Revised
33	C	2	3	0.83	-0.34	Discarded
34	B	0	0	0	0	Discarded
35	A	2	2	0.66	0	Discarded
36	C	1	1	0.33	0	Discarded
37	B	1	1	0.33	0	Discarded
38	D	2	1	0.55	0.33	Accepted
39	B	2	3	0.83	-0.34	Discarded
40	D	3	1	0.66	0.67	Accepted
41	A	2	0	0.33	0.66	Accepted
42	B	0	2	0.33	0	Discarded
43	B	3	2	0.83	0.34	Accepted
44	C	2	0	0.33	0.66	Accepted

From this table it can be see that of 44 test items, 33 (75%) are bad items in that they do not fulfill one or both of the requirements concerning the level of difficulty and power of discrimination. For example, test item 5, it meets the requirement of level of difficulty (0.66), but power of discrimination is in negative direction (-0.66) because more of the low group answer it correctly than those of the high group. This does not make sense because it is assumed that the number of the high group should be more than those of the low group in answering the test item correctly. If it happens the other way around, then the test item should be revised or discarded.

Whereas good items which meet the requirements of level of difficulty and power of discrimination amount to 11 items only (25%), they are: 2, 3, 4, 9, 10, 11, 21, 40, 41, 43, 44. In short, most items of this reading comprehension test do not function as a good test since only 25% of the total test items (11 items) meet the requirements as a good test.

b) Power of Discrimination

A test item is said to have power of discrimination if it can distinguish between the good and the poor students. If a test item is correctly answered by more of the low group than the high group, then that item has no power of discrimination

From the table before, it can been seen that there are 9 items which have negative direction, they are: 1, 5, 6, 16, 17, 29, 31, 33, and 39. This means that those items are correctly answered by more students of the low group than those of the high group. This seems ridiculous and it does not make sense because those students of the high group cannot answer that item correctly. Whereas those the low group can answer it. It is difficult to find out how it happens but that is the fact.

In addition, as many as 16 items have zero power of discrimination, they are item 7, 8, 12, 13, 14, 15, 18, 23, 24, 25, 30, 34, 35, 36, 37, 42. There are reasons why this can happen: (a) the number of low group and high group is the same in choosing the correct answer, for example, 3 students of the high group and 3 students of the low group, (b) some students of the low group answer the item correctly but none of the high group, (c) both the low group and high group do not choose the correct answer, or they both choose the wrong answer, but other students who do not belong to the group answer it correctly. Whatever it is, these 16 test

items are not acceptable because they have no power of discrimination which means they fail to distinguish between the high group and low group. The number of the high group and low group are the same in answering test items 7, 8, 12, 13, 14, 15, 18, 23, 24, 25, 30, 34, 35, 36 and 37. Even worse, for item 25, no one of both high group and low group answer it correctly. For item 42, two students of the low group answer it correctly, but none of the high group. Consequently, all these sixteen test items need revising as they do not meet the requirement of both level of difficulty and power discrimination. Theoretically, more of the high group should answer each item correctly than those of the low group. But if the two groups cannot answer one item correctly, there must be something wrong with the item which needs examining to find out what is wrong. It can be the distracters or the stem of the item is not very clearly stated. In case of item 42, in which two students of the low group answer it correctly and no students of the high group, that becomes questionable as such a thing rarely happens in a test.

Below are the tables showing the result of level of difficulty and power of discrimination.

Table 1. Level of Difficulty

Too difficult: 0.00 – 0.30	Good: 0.31 – 0.91	Too easy: 0.92 – 1.00
Item: 19, 20, 21, 22, 25, 27, 32, 34	Item: 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 16, 17, 18, 23, 24, 26, 28, 29, 30, 31, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44	Item: 7. 12, 15
8 items (8.18%)	33 items (75%)	3 items (6.82%)

The table shows that 75% of the total test have good level of difficulty, 6.82% is too easy, and 8.18% is too difficult for level of difficulty.

Table 2. Power of Discrimination

No/Poor Discrimination Power: 0.00 – 0.30	Average Discrimination power: 0.40 – 0.70	Good Discrimination Power 0.80 – 1.00
1, 3, 4, 5, 6, 7, 8, 9, 11 - 20, 22, 23 - 39, 42,43	2, 10, 21, 40, 41, 44	
38 items (83.36%)	6 items (13.64%)	

This table shows that only 13,64% of the total test have average power of discrimination, and another 86.36% have no or poor power of discrimination. Surprisingly, no items has good power of discrimination since there is none of those test items that reaches the coefficient of 0.80 and above.

In general, this reading comprehension test has poor discrimination power which means that the test items cannot distinguish between good students and poor students. That is why it is found out that one test item cannot be answered by all the test takers. On the other hand, one test item can be correctly answered by two students of the low group but none of the high group. To conclude, most test items need revising or discarded.

c) Effectiveness of Distracters

The multiple-choice item, including reading comprehension and vocabulary test, consists of (1) a stem or lead, which is either a direct question or incomplete statement, and (2) two or more choices responses of which one is the correct answer and the others are distracters, that is the incorrect responses (Harris, 1969: 7).The stem or lead must be clear and it is not recommended if it is only a word or two.

In order to know how good the distracters of this reading comprehension test are, below is presented the distracters picked up by the test takers, including the high and low group. One good characteristic of distracters is that each should be picked up by the test takers.

Item	A	B	C	D	Total
1	1	-	1	15 *	17
2	1	4	9 *	2	16
3	1	-	1	16 *	18
4	5	7 *	1	5	18
5	2	7	-	9 *	18
6	1	11 *	5	1	18
7	-	17 *	1	-	18
8	4	12 *	-	1	17
9	-	-	2	16 *	18
10	3	9 *	2	2	16
11	4	1	8 *	1	14
12	3	10 *	5	-	18
13	5	8 *	-	5	18
14	3	2	7 *	4	16
15	16 *	-	-	-	16
16	1	3	9 *	1	14
17	17 *	-	1	-	18
18	4	2	2	8 *	16
19	5	3	4 *	4	16
20	2	2 *	8	5	17
21	6	6	2 *	2	16
22	11	3 *	1	1	16
23	-	9 *	3	-	12
24	7	6 *	1	4	18
25	16	-	1	- *	17
26	3	4	6 *	3	16
27	4	7	4 *	1	16
28	4	8 *	2	-	14
29	4 *	4	7	-	15
30	4	3	1	7 *	15
31	8 *	1	4	2	15
32	7	4 *	1	4	16
33	2	1	10 *	1	14
34	2	- *	12	12	16
35	10 *	2	-	3	15
36	4	5	6 *	-	15
37	5	3 *	1	5	14

38	5	-	2	10 *	17
39	4	8 *	-	2	14
40	2	4	4	6 *	16
41	3 *	5	4	1	13
42	7	2 *	6	1	16
43	1	12 *	4	-	17
44	5	3	4 *	4	16

*= the correct answer

From the table, it can be seen which distracter of each item is not chosen by the test takers. Besides, some students do not choose any distracter of an item so that the number of the students vary accordingly. If the total number of the students are 16 or 14, it means there are two or four students not choosing any distracters. In other words, they abstain from choosing. There are 20 items (45.45%) the distracter of which is not chosen by the test takers, they are items:

1. B	8. C	15. B, C and D	28. D	36. D
3. B	9. A and B	17. B and D	29. D	39/ B
5. C	12. D	23. A and D	34. B*	39. C
7. A and D	13. C	25. B and D*	35. C	43. D

Of the items above, there are two items (4.54%), 25 and 34 the correct answer of which is not chosen by the test takers. It means no students know the correct answer of the two items, including those of the high and low group. In addition to that, there are six items in which two or three distracters are not chosen by the test takers. It is obvious that these 20 items need revising.

E. Conclusions and Suggestions

After having analyzed all those items of the reading comprehension test, the writer comes to the conclusions as the following:

(1) Despite the fact that this is standardized test or commercials one, all the requirements of a good test are not yet fulfilled. It is found out that 33 items (75%) do not meet the requirements of a good test and only 11 items (25%) meet both level of difficulty and power of discrimination.

(2) There are 16 items (36.36%) having no or zero power of discrimination because the number of both groups are the same in choosing the correct answer. Besides, there are 9 items with power of discrimination in negative direction because the number of the low group is bigger than the high group in choosing the correct answer.

(3) Twenty items (45.45%) of the total test items have one, two or three distracters which are not chosen by the test takers. Six of them with two or three distracters are not chosen. The correct answer of two items, 25 and 34, is not chosen by the test takers.

(4) Some items with one distracter not chosen still have adequate level of difficulty and power of discrimination because the number of the high group is bigger than the low group in choosing the correct answer.

(5) Some items (16, 31 and 33) have all the distracters chosen by the test takers although the number of the low group is bigger than the high group in choosing the correct answer. As a result, they have power of discrimination in negative direction but good level of difficulty.

(6) The difference between the high group and the low group is not very significant because there found only one score difference between the two. For example, the high group scores 28, and the low group 27. This difference is not very significant and that makes the power of discrimination in negative direction because the number of the low group is bigger than the high group in choosing the correct answer. This should not happen if the difference is significant.

As suggestions, the following is recommended:

(1) It is recommended that the teachers make item analysis of reading comprehension test before administering it in the classroom despite the fact that it is a standardized one. It may have too high vocabulary or other linguistics features which is not yet reached by the students. Therefore, the level of difficulty of the reading comprehension test should match the cognitive level of the students.

(2) The teachers who teach reading comprehension in the classroom should try to make their own reading comprehension test, often called "a teacher-made test" rather than giving the standardized test without being modified. The standardized one might be above the cognitive level of the students which can result in unsatisfactory outcome.

(3) Doing test item analysis and revising or changing bad distracters of the reading comprehension test will provide the teachers with new experience concerning reading comprehension test. In the long run, they will be familiar with this, and able to select one which suits the need of their students.

BIBLIOGRAPHY

Alyousef, S.2005. "Teaching Reading Comprehension to ESL/EFL Learners", The Reading Matrix Journal, 5(2) 144-150

Ardhana, Wayan, 1987. Bacaan Plihan Dalam Metode Penelitian Pendidikan. Jakarta: Depdikbud, Dikti.

Ary, Donald. et.all 1982. Introduction to Research in Education. New York: Holt Rineheart and Winston.

Cameron,L.2001. Teaching Language to Young Learners. Cambridge: Cambridge University Press.

Copperud, Carol. 1979. The Test Design Handbook. Englewood Cliffs: Educational Publicationl, Inc.

Gronlund, Norman E. 1985. Measurment and Evaluatikon in Teaching. New York: Macmillan Publishing Company.

Harris, David P. 1977. Testing English as a Second Language. New Delhi: Tata McGraw-Hill Publishing Company.

Madsen, Harold S. 1983. Technique in Testing. Oxford: Oxford University Press

Pebriawan, I. 2015. "The Correlation Between VocabularyMastery and Students' Reading Comprehension" Journal FKIP UNILA, 4(7), 123-144

Saleemi, Anjum P. 1988. "Language Testing: Some Fundamental Aspects", English Teaching Forum, Vol. XXVI, January 1988

Tuckman, Bruce W. 1975, Measuring Educational Outcomes: Fundamentals of Testing. New York: Harcourt Brace Jovanovich, Inc.

Valette, R.M. 1977. Modern Language Testing. New York: Harcourt Brace Jovanovich,Inc.