SCOPE
Journal of English Language Teaching

Article

Item Analysis of Reading Comprehension Test: A Study of Test Scores Interpretation

Risna Saswati

English Department, STBA LIA, Jalan Pengadegan Timur Raya No.3 Jakarta, Indonesia

CORRESPONDING AUTHOR(S):

E-mail: risnasaswati@gmail.com

A B S T R A C T

This study sheds lights on the effectiveness of multiple choice of reading comprehension test items in measuring learners' reading comprehension competence. This study applies descriptive quantitative method of research that uses statistical analysis to measure the effectiveness of the individual test item. The study uses the theories of multiple-choice test type and Item Analysis (IA). The data were taken from multiple choice test type of reading comprehension test given to 51 freshman students attending Literal Reading class. The test had 31 numbers comprising the questions of Text A, B and C. The reading comprehension test items were measured by the procedure of Item Analysis using the Item Facility, Item Discrimination and Distractor Efficiency. The result of this study is there are 13 items in reading comprehension test which are in need to be reviewed by the teachers. It is needed to be revised since the test is intended to be put in the question bank and be used repeatedly for the future use.

INTRODUCTION

Teaching involves assessment which is conducted to measure learners' ability and performance as a part of a process of teaching and learning. Teachers give learners assessment in the form of practices, tasks, projects, or other activities. Brown (2014) defines that assessment is ongoing process in teaching and learning that involves methodological techniques. It can be incidental or intended. Testing is a subset of assessment. They are prepared beforehand and are based on the school program regarding time and curriculum. Additionally, it is conducted after the learners master the units of the lesson. The learners know that they are measured and evaluated. They are aware about it. To conduct assessment, an instrument is needed. That is a test. Test is a method used to measure learner's ability and performance in a given domain. The key words for this definition are a method,

measure, learner's ability and performance, and given domain.

Brown (2014) describes those key words in details. A test is a method which means it is an instrument—a set of techniques, procedures, or items—that needs performance of the test takers to qualify the test itself. A test is to measure. It is a process of quantifying a test-taker's performance using procedures. A test is constructed based on the purposes. A test can act as a placement test which is intended to place a learner into a particular class. It could be diagnostic as well. The other key word is learner's ability and performance. A test measures learner's ability. A test should test learners based on their ability and it measures performance. The result implies the test taker's ability or competence. The given domain is related to test's purposes. A writing test, for example, should measure learner's performance in how they write a

good paragraph. The measurement comprises the content of the paragraph, punctuation, content, diction and organization of the paragraph.

Referring to test as an instrument, Brown adds that test can give accurate information about learner's ability and performance in a given domain. Therefore, constructing a good test is not an easy task to do. A test should fulfill the test criteria. It is valid, reliable and practical. The multiple-choice test is one of the types that teachers or test makers choose to measure the learners' competence. Besides it is easy to score, it is can be scored objectively. It has fair results. However, there are disadvantages when teachers employ this test to learners. Some of them are this type of test cannot measure performance since it adopts the recognition method of making test, it is not easy to construct and it can facilitate cheating and guessing. Those disadvantages can result negative backwash. Although it has disadvantages, the test makers or teachers choose this type of test. They sometimes use this test repeatedly. The question is how the test remains secure and can be used repeatedly. Bailey (1998) states that the teachers or test makers are asked to analyze the test by following the procedures. Therefore, the test can still work well and provide the valuable information for the test makers and teachers. The procedures are named by Item Analysis (IA, henceforth).

Additionally, teachers or test makers can analyze the test results by doing item analysis to qualify the test usefulness. Item analysis is a process to examine student's response to a test item in an objective test. It is to test the item whether it is qualified or not to measure what it is supposed to measure. It helps test maker to identify the item's difficulty and discrimination. The ineffective items can be discarded or revised for the future use. Item Analysis (IA, henceforth) is to qualify the test item which comprises Item Difficulty, Item Discrimination and Discrimination Efficiency.

One of the studies related to Multiple-choice test using the procedures of IA was conducted by Mehta and Mukhasi (2014). They investigated the quality of Multiple-Choice questions to create the qualified question bank for future use. There were 50 questions undertaken by a hundred first year students. It was conducted in the Department of Anatomy. Regarding the procedures, this study employed IA using Item Difficulty, Item Discrimination, and Distractor Efficiency. The results of this study revealed that the Item Difficulty reached 62%, Discrimination Efficiency was 53% consisting of non-functional distractors and 38% had no responses. This study inferred that some numbers or items should be reviewed to result qualified test items and could be used for future use.

Another one was conducted by Muhson, et. al., (2017), it shed lights on the use of a machine named by AnBuso to analyze the multiple-choice and essay test. The machine could be used by teachers to analyze the test they made. The test analyzed could be stored in the question bank for future use. AnBuso was developed based on CTT (Classical Test Theory). The method employed was Research and Development (R&D) using the procedure of designing, developing and validating and trying the product. It involved 65 respondents comprising teachers and supervisors in Yogyakarta. The result of this study was the willingness for teachers to do IA using Anbuso was low which was 57%. There were 11% of the respondents analyzed the test, 2% of the respondents often did the IA, and 12% of them never did IA. The teachers could use the analyzed test for future use when the test was used repeatedly.

The other study conducted by Burud and Nagandla (2019) investigated the impact of distractors of the multiple-choice test. It was to evaluate the quality of the test items using IA analysis by using Item Difficulty, Item Discrimination and Discrimination Efficiency and assessment of functional and non-functional distractors. The data were taken from summative test undertaken by 113 test takers. There were 360 distractors and chosen 85 only by the respondents. The result was majority of the items showed excellent difficulty index (50.4%) and fair discrimination (37%). Additionally, 13% of the data was found had no functional distractors. The distractors efficiency gave the information to the teachers or test makers on the quality of test items.

The previous studies shed lights on the Item Analysis using the discrimination efficiency procedure (Burud and Nagandla, 2019), AnBuso as the machine to be used by teachers and supervisors to analyze Item Analysis of the multiple-choice test and essay test (Muhson, et. al., 2017) and Item Analysis of multiple-choice questions (Mehta and Mukhasi, 2014). This study was different from those discussed previously in terms of the data and the research methodology. The results presented in this study were also different since it used the number of items in the test. Those numbers stated should be replaced and revised since the test will be used repeatedly for the next semester.

Therefore, this study investigates whether the test items of reading comprehension tests used in Literal Reading supplementary test is effective or not for the future use. Additionally, it is to study the items regarding its discrimination efficiency (IE, henceforth). The IA describes the statistical analysis which allows the measurement the effectiveness of individual test items. IA

can draw clear description about what remedial work can be given to students. Regarding performance, it is to compile future test. Teachers have tendency to use the test again. Therefore, it is also beneficial to identify the test items based on IA. Information supplied in this study will equip teachers to develop, analyze, select and refine the test items more appropriate to be used again for testing the students. Implicitly, the items that are considered discarded is recommended to be revised or changed using the new item. Therefore, the revised test is supposed to be much better in measuring students' competence in reading comprehension. Based on the research gaps, this study attempts to find out how well the items work well with particular group of students. Regarding test items, the IA results are to inform which items are to be removed or revised based on Item Distractor Efficiency Analysis.

This study describes the segments of:

a. Multiple Choice Test

The question raised is why teachers use multiple choice items so often. The reasons noted by Bailey (1998) are:

1. Multiple choice tests are easy to score.
   It is not difficult to score Multiple choice test since they consist of one correct answer in the form of A, B, C or D. It can be scored by a person or more than one person who do not necessarily have the same educational background with the test takers.
2. They can be scored objectively. It seems fair and reliable to use this type of objective tests compared to other types of subjective tests.
3. They are tests and acceptable by convention.
4. They can reduce the guessing compared to true false items which have two choices only. If it is compared to true false items, it is believed to reduce guessing since they have one correct answer and three distractors or more.

Besides the pluses, multiple choices have drawbacks to be chosen as type of test (Heaton, 1988). Those are:

1. It is not easy to construct.
   It is not easy to find the options and one correct answer. There is a rule to obey related to choose the options which consist of a correct answer and the others are distractors. As the name suggest, the distractors function to distract the correct answer. It should be parallel in the level of class of words and the length. When one option is the level of word, the other options should be words and their classes of words are the same.
2. It takes time to construct this test.
   Test maker should create one problem and 3 or 4 or 5 choices and one correct answer. It takes time to find the choices. Additionally, it must be tested to colleagues to result the validity of the test.
3. It has negative washback.

It is believed that this type of test facilitates guessing. Test takers can guess the answers without considering much about the correct answer.
4. It has the possibility that the results cannot provide accurate measurements. The guessing done by learners can trigger bias in test results. Learners have the opportunity to cheat from their friends.

Hughes (1989) notes that multiple choice items bring a discussion:

1. It is recognition test.
2. Guessing of the items can bring effect to the test scores.
3. The technique restricts what to be tested.
4. It is difficult to write down a successful item.
5. Backwash is probably harmful.
6. Cheating may be facilitated.

Multiple choice consists of stem with three or four or five options. It has one correct answer and the other options which are not the correct answers play as distractors. Based on the consideration teachers are advisable to do item analysis to qualify the test.

b. Item Analysis

Item analysis is the systematic evaluation of the effectiveness of the items of the test. It is conducted to select the best items that work with particular group of learners. IA when used for Norm Referenced Tests (NRTs) applies item format analysis, Item Facility (IF) analysis, Item Discrimination (ID) analysis and Distractor Efficiency analysis. IF comprises of item difficulty and item discrimination (Heaton, 1988; Bailey, 1998; Brown and Abeywickrama, 2010). There is an addition in calculating whether or not the items are changed. It is categorized as discrimination efficiency (Brown, 2014) or Discrimination Analysis (Bailey, 1998).

1. Item Difficulty
   It is an index of how easy an item for an individual to take it. The index represents the test takers who get right answers. It presents the interesting information relative to the entire group of test takes of each individual item.
2. Item Discrimination
   The index represents in which the item is too easy for the test takers and too difficult for them. If it is found as an evidence the item should be discarded from the tests.
3. Discrimination Efficiency or Discrimination Analysis
   It is to see how individual distractor functions. It is based on the considerations that some distractors are not distracting at all. It serves no purpose. It can be found out using this discrimination efficiency. To qualify the test items, some distractors are eliminated.

Those techniques are used for NRTs which is defined as a test that measures how the performance of particular test takers or group compares with the performance of another test taker whose scores are given as norms. The test scores are therefore interpreted with reference to the scores of other test takers rather than to the criteria that are agreed.

METHOD

The objectives of the study are to shed lights on whether or not the test items in reading comprehension tests are effective based on the measurement of IA which comprises the item difficulty, item discrimination and distractor efficiency. This study is to implement the research methodology as it is written following.

It is to apply the descriptive quantitative method to describe the qualification of the refined test. The approach employed is the Classical Test Theory (CTT) that uses statistics to measure the effectiveness of the individual test items. The calculation procedures of Item Analysis were employed using the Item Difficulty, Item Discrimination and Discrimination Efficiency.

The data are items taken from multiple choice test of reading comprehension test used as supplementary material for first semester of Literal Reading class at one of the universities in Jakarta. The test as the instrument was taken from the text book used for the reading class as the supplementary material. The multiple-choice test comprises three texts, which consists of 31 numbers. It is given as the assessment to measure their reading comprehension undertaken by 51 test takers.

The data are taken from multiple choice items of reading comprehension test. There are 31 items undertaken by 51 learners. to be calculated using Item Analysis which consists of Item Facility, Item Discrimination and Distractor Efficiency. To measure IA, the data are calculated based on the formulas.

It applies statistical measurement. The study is to calculate Item Facility first to find out how many students get the correct answer. The item discrimination is applied after the IF is counted which is to find out the degree to which an item separates the students perform well from the ones who perform poorly. Then, the distractor efficiency is calculated to shed lights on the selection of the items to be used in a revised version of a test. It is to find out as well the distractors function well or not.

RESULTS AND DISCUSSION

This part presents the results and discussion which comprises of the Item difficulty, Item Discrimination and Discrimination Efficiency. Item Analysis consists of Item Facility (IF) which is a statistical index that is used to examine the percentage of students who choose the correct answers. To calculate the IF index, the number of students answer the items correctly is divided by all students taking the test. The table following reveals how many students answer the items correctly.

Table 1 : Item Facility Index

No.	Option A	Option B	Option C	Total Answers	IF Index
TEXT 1					
1	17	6	28	51	0.55
2	0	32	19	51	0.37
3	3	43	5	51	0.84
4	47	2	2	51	0.92
5	17	2	32	51	0.63
6	3	3	45	51	0.88
7	16	1	34	51	0.31
8	10	34	5	51	0.67
9	15	21	15	51	0.41
10	28	4	19	51	0.55
TEXT 2					
11	22	28	1	51	0.55
12	6	31	14	51	0.61
13	24	22	5	51	0.43
14	13	12	26	51	0.51
15	20	20	11	51	0.39
16	10	9	32	51	0.63
17	4	20	27	51	0.53
18	30	13	8	51	0.61
19	0	13	38	51	0.75
20	5	28	18	51	0.55
TEXT 3					
21	18	30	3	51	0.59
22	34	10	7	51	0.67
23	32	11	8	51	0.63
24	18	3	30	51	0.59
25	10	3	38	51	0.75
26	25	18	8	51	0.35
27	34	6	11	51	0.69
28	8	13	30	51	0.59
29	5	24	22	51	0.47
30	13	4	34	51	0.67
31	26	23	2	51	0.51

The test consisted of 31 numbers which were given to 51 test takers. Text 1 consisted of 10 test items, text 2 had 10 test items and text 3 consisted of 11 numbers. All the test items were taken from the text book. The Multiple-choice items had three choices in which there were two distractors and one correct answer. The table gave the evidences about which options were answered correctly by the test takers. Item number 1 had 17 correct answers for option A, 6 correct answers for option B, and 28 correct answers for option C. There were 55% of the test takers answered this item correctly which meant the item was average regarding the level of difficulty. The implication for the test maker was that the test was not necessary to be replaced. Compared to test item number 4, it should be replaced by the new item since it was too

easy for the test takers. It revealed that the numbers of test takers answered it correctly was 92%. When the item has IF index 0,35 or 35%, it was to be replaced or revised since only 35% answered the item correctly. The item was considered too difficult for the test takers to answer.

The implication for the test maker was that there were items that were considered very difficult. Those were items numbers 2 and 13 which were answered correctly by only 37% and 35% of all test takers. Besides, there were items considered very easy to answer since it reached 92% and 88% which meant there were 43 and 45 students out of 51 students answered those items correctly. The items considered to be revised and improved were items number 3, 4, 6, 19, and 26.

The next calculation is the item discrimination index in which it is to separate the test takers into two groups which are the upper group and lower groups. The upper group consists of students who perform well and the lower group consists of students who perform poorly. The reason for identifying those is the item discrimination is to contrast the performance of the two groups. The data are sorted into group of students based on the scores that they have which is listed based on the highest to the lowest and divided into two groups: upper and lower groups. To calculate the ID index, the IF of the upper group is subtracted from the IF of the lower group. Table 2 showed the results.

Table 2 : Item Discrimination

Item Discrimination																															
ITEMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
CORRECT ANSWER	28	19	43	47	32	45	16	34	21	28	28	31	22	26	20	32	27	31	38	28	30	34	33	30	38	18	34	32	23	34	26
CA UPPER GROUP	12	11	26	26	19	26	8	23	10	16	17	21	17	17	10	21	11	16	22	18	20	22	18	16	25	13	22	19	15	22	14
CA LOWER GROUP	16	8	17	21	13	19	8	11	11	12	11	10	6	9	10	11	16	15	16	10	10	12	15	14	13	5	12	13	8	12	12
IF	0.55	0.37	0.84	0.92	0.63	0.88	0.31	0.67	0.41	0.55	0.55	0.61	0.43	0.51	0.39	0.63	0.53	0.61	0.75	0.55	0.59	0.67	0.65	0.59	0.75	0.35	0.67	0.63	0.45	0.67	0.51
IF UPPER	0.48	0.44	1.04	1.04	0.76	1.04	0.32	0.92	0.4	0.64	0.68	0.84	0.68	0.68	0.4	0.84	0.44	0.64	0.88	0.72	0.8	0.88	0.72	0.64	1	0.52	0.88	0.76	0.6	0.88	0.56
IF LOWER	0.61	0.31	0.65	0.81	0.50	0.73	0.31	0.42	0.42	0.46	0.42	0.38	0.23	0.35	0.38	0.42	0.62	0.58	0.62	0.38	0.38	0.46	0.58	0.54	0.50	0.19	0.46	0.50	0.31	0.46	0.46
ID	-0.13	0.13	0.39	0.23	0.26	0.31	0.01	0.50	-0.02	0.18	0.26	0.46	0.45	0.33	0.02	0.42	-0.18	0.06	0.26	0.34	0.42	0.42	0.14	0.10	0.50	0.33	0.42	0.26	0.29	0.42	0.10

The table showed the number of correct answers. Item number 1 had 28 correct answers (CA) which was divided by 12 correct answers gained from the upper group and 16 correct answers taken from the lower group. To get the group division, it was seen from the number of test takers, 51 test takers, divided into two. It was 25 test takers for the upper group and 26 test takers for the lower group. Then, the correct answers for each item were calculated. The IF index for each group was calculated as similar as the IF for the whole group. The Item Discrimination (ID) index was gained from the subtraction from the IF upper and lower group of the test takers. Whether or not the item was replaced or revised, the test maker should check the guidelines based on the score of the Item Discrimination (ID) index.

The guidelines to categorize the items is taken from Ebel quoted by Brown (2010) that should not be used fixed rules. However, it aids the teachers and test makers in making decisions about which items are to keep and to discard. The guidelines suggested are:

.40 and up   very good items

.30 to .39   reasonably good but possibly subject to improvement

.20 to .29   marginal items, usually needing and being subject to improvement

Below .19   poor items, to be rejected or improved by revision

Based on it, there were items that are categorized as poor items based on ID indexes. Items number 1, 2, 7, 9, 15, 16, 17, 18, 23, 24, and 31. Those items are decided to be revised and improved if teachers need to use the test again for different group of students. While selecting those items based on ID index, the test is under development status which means it is not appropriate to be used as a test. The development is to make the test more suitable to measure students' proficiency.

The last calculation was to find out the discrimination efficiency using the Discrimination Efficiency formula. It is to select the items to be used in a revised and improved version of a test. The statistical analysis reveals that different part of the item can help the test makers to make sure that the options of each item functions well. The following table shows the evidence about the test items' options. The previous calculation, IF and ID are not enough to select the best items. The job of revising and improving the test is not finished yet, especially for the multiple- choice items. Distractor Efficient Analysis is to examine the degree to which the distractors function well even for the students who do not know the answers. The distractor can function well to distract the other options. To calculate it, the test takers are divided into three groups: Upper, Middle and Lower group.

Table 3 : Distractor Efficient Analysis

| ITEMS | IF | ID | GROUP | OPTIONS | | | |
				A	B	C	TOTAL
1	0.55	0.13	Upper	0.40	0.35	0.25	1
			Middle	0.29	0.06	0.65	1
			Lower	0.24	0.12	0.65	1
2	0.37	0.13	Upper	0.00	0.53	0.47	1
			Middle	0.00	0.59	0.41	1
			Lower	0.00	0.71	0.29	1
3	0.84	0.39	Upper	0.18	0.59	0.24	1
			Middle	0.00	0.94	0.06	1
			Lower	0.18	0.59	0.24	1
4	0.92	0.23	Upper	1.00	0.00	0.00	1
			Middle	0.94	0.06	0.00	1
			Lower	082	0.06	0.12	1
5	0.63	0.26	Upper	0.29	0.00	0.71	1
			Middle	0.29	0.06	0.94	1
			Lower	0.41	0.06	0.71	1
6	0.88	0.31	Upper	0.00	0.00	0.71	1
			Middle	0.06	0.06	0.94	1
			Lower	0.18	0.18	0.71	1
7	0.31	0.01	Upper	0.41	0.00	0.10	1
			Middle	0.29	0.71	0.00	1
			Lower	0.18	0.06	0.52	1
8	0.67	0.5	Upper	0.00	1.00	0.00	1
			Middle	0.35	0.65	0.00	1
			Lower	0.41	0.35	0.24	1
9	0.41	0.02	Upper	0.47	0.35	0.18	1
			Middle	0.12	0.47	0.41	1
			Lower	0.29	0.35	0.35	1
10	0.55	0.18	Upper	0.59	0.12	0.29	1
			Middle	0.71	0.00	0.29	1
			Lower	0.35	0.12	0.53	1
11	0.55	0.26	Upper	0.35	0.65	0.00	1
			Middle	0.47	0.47	0.06	1
			Lower	0.53	0.47	0.00	1
12	0.61	0.46	Upper	0.00	0.94	0.06	1
			Middle	0.00	0.53	0.47	1
			Lower	0.35	0.35	0.29	1
13	0.43	0.45	Upper	0.24	0.65	0.12	1
			Middle	0.53	0.47	0.00	1
			Lower	0.59	0.24	0.18	1
14	0.51	0.33	Upper	0.71	0.00	0.29	1
			Middle	0.24	0.29	0.47	1

			Lower	0.41	0.24	0.35	1
15	0.39	0.02	Upper	0.35	0.29	0.35	1
			Middle	0.47	0.41	0.12	1
			Lower	0.35	0.47	0.18	1
16	0.63	0.42	Upper	0.00	0..06	0.94	1
			Middle	0.24	0.12	0.65	1
			Lower	0.24	0.29	0.47	1
17	0.53	0.18	Upper	0.06	0.53	0.41	1
			Middle	0.00	0.47	0.53	1
			Lower	0.12	0.24	0.65	1
18	0.61	0.18	Upper	0.71	0.18	0.12	1
			Middle	0.59	0.18	0.24	1
			Lower	0.53	0.35	0.12	1
19	0.75	0.26	Upper	0.00	0.06	0.94	1
			Middle	0.29	0.00	0.71	1
			Lower	0.00	0.41	0.59	1
20	0.55	0.34	Upper	0.00	0.71	0.29	1
			Middle	0.12	0.47	0.41	1
			Lower	0.18	0.47	0.35	1
21	0.59	0.42	Upper	0.18	0.76	0.06	1
			Middle	0.59	0.41	0.00	1
			Lower	0.35	0.53	0.12	1
22	0.67	0.42	Upper	0.94	0.06	0.00	1
			Middle	0.53	0.35	0.12	1
			Lower	0.53	0.18	0.29	1
23	0.65	0.14	Upper	0.71	0.06	0.00	1
			Middle	0.53	0.35	0.12	1
			Lower	0.53	0.18	0.29	1
24	0.59	0.1	Upper	0.18	0.00	0.82	1
			Middle	0.00	0.59	0.40	1
			Lower	0.47	0.18	0.53	1
25	0.75	0.5	Upper	0.00	0.00	1.00	1
			Middle	0.29	0.06	0.65	1
			Lower	0.41	0.12	0.47	1
26	0.35	0.33	Upper	0.18	0.71	0.12	1
			Middle	0.65	0.12	0.24	1
			Lower	0.53	0.24	0.24	1
27	0.67	0.42	Upper	1.00	0.00	0.00	1
			Middle	0.65	0.12	0.24	1
			Lower	0.35	0.24	0.24	1
28	0.63	0.26	Upper	0.18	0.06	0.76	1
			Middle	0.12	0.29	0.59	1
			Lower	0.24	0.24	0.41	1
29	0.45	0.29	Upper	0.00	0.65	0.35	1
			Middle	0.12	0.47	0.41	1
			Lower	0.18	0.18	0.06	1
30	0.67	0.42	Upper	0.67	0.06	0.88	1
			Middle	0.29	0.06	0.65	1
			Lower	0.35	0.18	0.47	1
31	0.51	0.10	Upper	0.53	0.41	0.06	1
			Middle	0.53	0.47	0.00	1

Lower	0.47	0.47	0.06	1

Based on the evidence, item number 2, option A, does not discriminate the high, middle and lower group of students as indicated by ID index. Additionally, it is answered only 37% by the students. Another item is number 4 in which option B and C are not attracted all the high group and attract only a small number of students in the middle and lower level. The other items are in need to be revised in terms of how these items function well to distract other options are item number 5 for option B, number 6 for option A and B, number 8 for all options, number 12 for option A and number 19 for option A. These items are needed to revised and improved by the test makers.

CONCLUSION

Item Analysis which consists of the techniques of IF, ID and Discrimination Efficient Analysis can be used as a tool to develop, analyze, select and refine test items. It is conducted on account of that teachers or test makers examine only the test scores of students which is not fair. It is to consider whether the test is measured what it is support to be measured. The test for measuring how well the students comprehend the text used as assessment is considered to be revised and improved to have a suitable version to test the students. Test makers and teachers should keep these statistical techniques as only a tool for improving the test items to result the revised and improved version that can test the students properly. Based on the results of Item Difficulty, the items should be reviewed are number 3, 4, 6, 19 and 26. Based on the calculation of the Item Discrimination, the numbers which were considered poor or to be rejected were number 1, 2, 7, 9, 15, 16, 17, 18, 23, 24, and 31. The Discrimination Efficiency viewed the numbers from the options or distractors. The distractors should be revised were the options on number 2, 4, 5, 6, 8, 12, 25 and 19. Therefore, the numbers should be reviewed by the teachers or test makers were number 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 15, 16, 17, 18, 19, 23, 24, 25, 26 and 31. There

were 10 numbers which were in need to be reviewed. When it is intended to be used repeatedly, the teachers or test makers are recommended to revise all the numbers stated.

REFERENCES

Bailey, K. M. (1998). Learning About Language Assessment, Dilemmas, Decisions, and Directions. USA: Heinle & Heinle Publishers.

Brown & Abeywicakrama, P. (2010). Language Assessment: Principles and Classroom Practices. New York: Pearson.

Brown, D. J. (2014). Testing in Language Programs: A Comprehensive Guide to English Language Assessment. US: JD Publisher.

Burud, I., & Nagandla, K. (2019). Impact of distractors in item analysis of multiple-choice questions. International Journal of Research of Medical Sciences, 7(4), 1136-1139.

Heaton, J. B. (1988). Writing English Language Test. London: Longman.

Hughes, A. (1989). Testing for Language Teachers. Cambridge: Cambridge University Press.

Mehta, G., & Mukhasi, V. (2014). Item analysis of multiple-choice questions. International Journal of Health and Sciences and Research, 7(4), 197-202.

Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of the practical item analysis program for Indonesian teachers. International Journal of Instruction, 10(2), 192-210.