ISSN 2808-2265

Language Studies Centre

Pattimura Excellence Journal of Language and Culture

Journal Home page: https://ojs3.unpatti.ac.id/index.php/pejlac
Journal Email: pejlac.unpatti@gmail.com

ITEM ANALYSIS OF AN ENGLISH SUMMATIVE TEST

Leni Amelia Suek

Nusa Cendana University, Indonesia

Article Information:

*Corresponding author
leni.suek@yahoo.com

Received March, 8 2021;
Revised April 10, 2021;
Accepted May 17, 2021;
Publication June 1, 2021

Abstract. While almost half of teachers' activities are assessing their students, they are not well-prepared with assessment literacy training. Hence, they are unable produce good tests to measure students' level of knowledge and skills. This study is aimed at analyzing item difficulty and item discrimination of a multiple-choice test made by an English teacher at a junior high school in Kupang. The instruments of this qualitative research were 50 test items and students' answer sheets. The results of this study indicated that the English summative test had poor item difficulty index and low item discrimination index. For the level of difficulty, it was found that 27 items (54%) were easy, 22 items (44%) were moderate, and 1 item (2%) was difficult. For item discrimination power, it was revealed that 5 items (10%) were excellent, 8 items (16%) were good, 8 items (16%) were satisfactory, 23 items (46%) were poor, and 6 items (12%) were negative. In addition, 12 items were acceptable, 20 items were unacceptable and 18 items needed revisions. In conclusion, this English summative test did not fulfill the criteria of a good test and could not measure students' true ability.

Keywords*: item difficulty, item discrimination, English test, criteria of a good test*

1. INTRODUCTION

Teachers have a set of responsibilities which is not only about preparing and teaching the lesson but also assessing students and evaluating the course. One of the ways to measure students' ability is by conducting a test (McNamara, 2000; Hughes, 1989). It is important for the teachers to make good test items because this can measure students' true ability, reflect the success of the lesson and also indicate low and high performers. This is also a part of assessment and evaluation because testing is a device to reflect the assessment process and the effectiveness of the lesson and teaching process (Fulcher & Davidson, 2007). Therefore, developing lessons is as important as evaluating the lessons.

Teachers, these days, are not well-prepared with assessment training. They focus on developing teaching method but overlook to equip themselves with assessment literacy which may lead to conducting poor test items. Almost half of teachers' activities are assessing their students, but they are not well-prepared to make a good assessment (Plake & Impara, 1996). They make the test to measure students' ability but it does not fulfill the criteria of a good test such as validity, reliability and practicality (Hughes, 1989). In addition, they often construct multiple choice tests with poor level of difficulty and low item discrimination. This means that they are unable to make valid and reliable test. A well-constructed classroom test will provide students with an opportunity to show their ability to recognize and to produce correct forms of the language.

Multiple choice-test is one types of tests that is frequently used by English teachers to measure their students' English knowledge and skills. Multiple-choice tests are receptive or selective kinds of tests (Toksoz & Ertunc, 2017). This test requires the test takers to select one of the options from the test item. This type of test has stem, options and distractors. The stem contains the statement or the question and the options are alternatives to be selected. The right option must be selected by eliminating the wrong answers or distractors (Brown, 2004; Hughes, 1989).

Teachers prefer multiple choice item because of its practicality. It does not consume much time to prepare and it is easy to administer. This test is also more reliable than any other types of tests because it is objective and scored consistently (Brown, 2004). However, this kind of test may have low discrimination power because it does not clearly discriminate high performers and low performers if the items are not well-designed. In this case, students might just guess the answers without reflecting their knowledge or skills. The quality of a multiple-choice test depends on the test items (Brown, 2004; Hughes, 1989; Kehoe, 19951). If teachers are unable to design the multiple-choice items well, the students might just guess the answer. In fact, guessing might affect the scores of the students (Brown, 2004; McNamara, 2000). One of the biggest challenges in designing multiple-choice items is writing successful items that fulfill the criteria of validity, item difficulty and item discrimination.

In order to write successful multiple-choice items, the test should have good item difficulty and high item discrimination (Haladyna, 2004; Henning, 1987). Item difficulty determines which items are difficult and which items are easy (Brown, 2004). It also determines whether the questions are trivial, difficult or impossible items (Bodner, 1980). While, item discrimination functions to differentiate higher and lower performers (Brown, 2004). An item with high discrimination means that good students can get it right while poor students will get it wrong (Toksoz & Ertunc, 2017).

The increasing trend of using multiple-choice items for testing students' ability and the fact that teachers could construct good tests encouraged the researcher to analyze the multiple-choice test items made by an English teacher. This study is aimed at finding out the item difficulty index and item discrimination index of a multiple-choice test made by and English teacher at SMP Negeri 1 Kupang. This research will shed a light on the tests constructed by the teachers, the quality of test, measurement of student's ability and also the implications for further test construction.

2. RESEARCH METHODOLOGY

Documentation and descriptive methods with quantitative approach were used in this study. First, the multiple-choice test that has was constructed by an English teacher at the ninth grade of SMP Negeri 1 Kupang Tengah was analyzed. A quantitative approach was used to measure the difficulty level and discrimination power of the test. The subject of this study was the English teacher and the students at the ninth grade of SMP Negeri 1 Kupang Tengah academic year 2019/2020. This school was selected because it has been accredited A since a few years ago which means that it is a good school and it is expected that the teachers are able to construct good tests. The instrument used in this study was 50 multiple-choice test questions that was constructed by the English teacher and 36 students' answer sheets that were corrected or scored by the teacher.

In order to collect data, a series of procedure was taken. First, the researcher went to the school and met the English teacher to get the test items and students' answers sheets and scores. By her permission, the English summative test was analyzed by examining the students' answer sheets and scores, computing the difficulty levels and the discrimination power of all items, revealing good and poor test items, discussing the findings and making conclusions. Arifin (2012:251) said that there are several steps to analyze test items. First, score all of the students' answer sheets. Then, the scores are recorded ranging from highest to the lowest. Next, 27% of higher performers and 27% of low performers are grouped. While the medium performers are put aside. Finally, students' answers are analyzed. An item analysis reveals three things including: how difficult each item is, whether or not the question discriminates or tells different between high and low students, and which distractors are working as they should.

In constructing a test, teachers should consider the difficulty level of the items. To claim that the item is easy or difficult, an analyzes of students' answer sheets has to be done. A test item is too easy when more than 90% of the students answered it correctly. An item is too difficult when less than 30% of the students answered it correctly. According to Arifin (2012: 266), the steps to find out the level of difficulty begin by tabulating students' answer sheets from the highest scores (high group) to the lowest scores (low group). Then, 27% of the answer sheet from the high group and 27% of the answer sheet from the low group was taken while the remaining 46% was set aside. Finally, all the data was presented in the table to find out the answers from each student both from the high and low groups. The score for the correct answer is 1, and the wrong answer will get 0.

The level of difficulty is calculated as follows (Arifin, 2012: 266):

$$\boxed{\mathrm{DL} = \frac{WL + W\mathrm{H}}{nL + n\mathrm{H}} x100\%}$$

DL = difficulty level
WL = total students who answered incorrectly from the lower group
WH = total students who answered incorrectly from the higher group
nL = total students in lower groups
nH = total students in higher groups

The criteria for interpreting the difficulty level that was claimed by Arifin (2012: 270) ranging from easy to difficult as follows:

If DL is 27%, the test item is easy.
If DL is between 28% and 72%, the test item is moderate.
If DL is 73%, the test item is difficult.

Discrimination power discriminates high performers from low performers. According to Arifin (2012: 274), there are several procedures in analyzing the discrimination power including tabulating students' answer sheet, counting the number of students who got the item wrong in the low group (WL) and counting the number of students who got the item wrong in the high group (WH), subtracting WL to WH, and calculating the discrimination power of each question.

The discrimination power is calculated as follows (Arifin, 2012: 266; Hoha, 2001:147)

$$\boxed{\mathrm{Dp} = \frac{WL - WH}{n}}$$

Dp = discrimination power
WL = total students who answered incorrectly from the lower group
WH = total students who answered incorrectly from the higher group
n = total students

The criteria for interpreting the discrimination power that was claimed by Arifin (2012: 270) ranging from poor to excellent as follows:

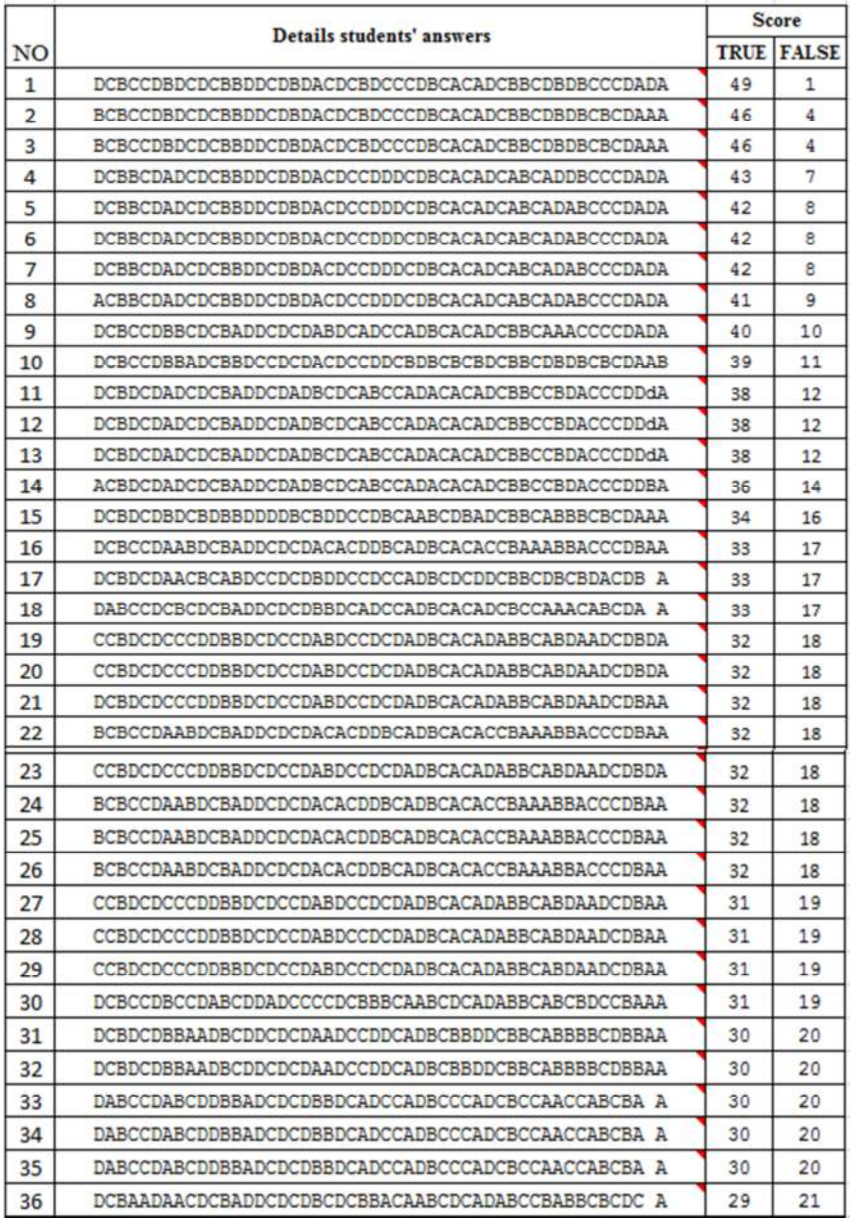
0.00 – 0.20	Poor
0.20 – 0.40	Satisfactory
0.40 – 0.70	Good
0.70 – 1.00	Excellent
Minus	Negative

3. FINDINGS AND DISCUSSION

Arikunto (2008: 206-207) stated it is important to analyze the questions that have been made to find which items are excellent and which items are poor. According to Sudijono (2009: 376-378) well-constructed items should be used in the test, while poor items should be revised or changed. In this study, the English summative test was analyzed to find out the level of difficulty and discrimination power. The test was made at the end of the semester by the teacher to assess the materials that have been taught. It was also used to prepare the students for the national examination. Summative test according to Djamarah (2005; 253) is an assessment that is carried out at the end of each teaching of a program or a certain number of learning units.

The findings are presented based on on the steps in analyzing the tests. The first step in doing the item analysis is to score all of the students' answer sheets. Students' scores are presented in the following table.

Table 1. Students' scores

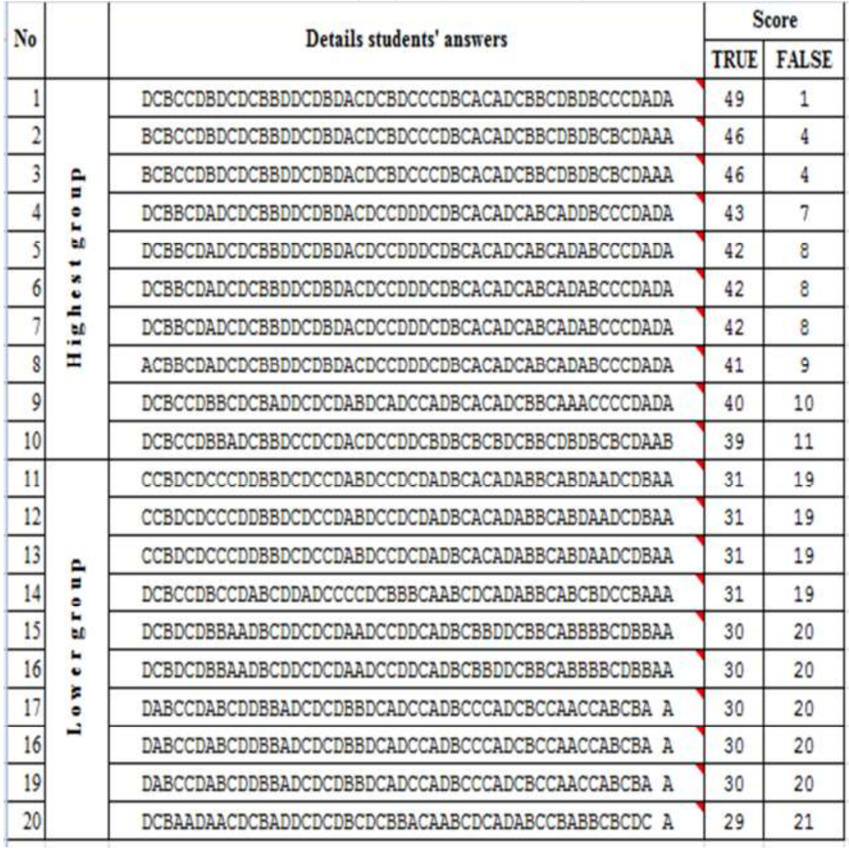


NO	Details students' answers	Score	
		TRUE	FALSE
1	DCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCCCDADA	49	1
2	BCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCBCDAAA	46	4
3	BCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCBCDAAA	46	4
4	DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADDBCCCDADA	43	7
5	DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
6	DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
7	DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
8	ACBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	41	9
9	DCBCCDBBCDCBADDCDCDABDCADCCADBCACADCBBCAAACCCCDADA	40	10
10	DCBCCDBBADCBBDCCDCDACDCCDDCBDBCBCBDCBBCDBDBCBCDAAB	39	11
11	DCBDCDADCDCBADDCDADBCDCABCCADACACADCBBCCBDACCCDDdA	38	12
12	DCBDCDADCDCBADDCDADBCDCABCCADACACADCBBCCBDACCCDDdA	38	12
13	DCBDCDADCDCBADDCDADBCDCABCCADACACADCBBCCBDACCCDDdA	38	12
14	ACBDCDADCDCBADDCDADBCDCABCCADACACADCBBCCBDACCCDDBA	36	14
15	DCBDCDBDCBDBBDDDDBCBDDCCDBCAABCDBADCBBCABBBCBCDAAA	34	16
16	DCBCCDAABDCBADDCDCDACACDDBCADBCACACCBAAABBACCCDBAA	33	17
17	DCBDCDAACBCABDCCDCDBDDCCDCCADBCDCDDCBBCDBCBDACDB A	33	17
18	DABCCDCBCDCBADDCDCDBBDCADCCADBCACADCBCCAAACABCDA A	33	17
19	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBDA	32	18
20	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBDA	32	18
21	DCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	32	18
22	BCBCCDAABDCBADDCDCDACACDDBCADBCACACCBAAABBACCCDBAA	32	18
23	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBDA	32	18
24	BCBCCDAABDCBADDCDCDACACDDBCADBCACACCBAAABBACCCDBAA	32	18
25	BCBCCDAABDCBADDCDCDACACDDBCADBCACACCBAAABBACCCDBAA	32	18
26	BCBCCDAABDCBADDCDCDACACDDBCADBCACACCBAAABBACCCDBAA	32	18
27	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
28	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
29	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
30	DCBCCDBCCDABCDDADCCCCDCBBBCAABCDCADABBCABCBDCCBAAA	31	19
31	DCBDCDBBAADBCDDCDCDAADCCDDCADBCBBDDCBBCABBBBCDBBAA	30	20
32	DCBDCDBBAADBCDDCDCDAADCCDDCADBCBBDDCBBCABBBBCDBBAA	30	20
33	DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
34	DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
35	DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
36	DCBAADAACDCBADDCDCDBCDCBBACAABCDCADABCCBABBCBCDC A	29	21

Table 1 indicated scores of all 36 students who took the test. The student with the highest score answered 49 out of 50 questions correctly, while the lowest one answered 29 out of 50 questions correctly. Students with the lowest score answered 21 questions incorrectly and the student with the highest score answered 1 question incorrectly.

The second step is deciding 27% of high group and 27% low group. According to Arifin (2012: 266), the steps to find the level of difficulty items begin by tabulating students' answer sheets from the highest score (high group) to the lowest score (low group). Then, 27% of the answer sheet from the high group and 27% of the answer sheet from the low group was taken while the remaining 46% was set aside. Top 10 students were grouped as hig performers (27% x 36 students) and 10 students with the lowest score were grouped as low performers (27% x 36 students).

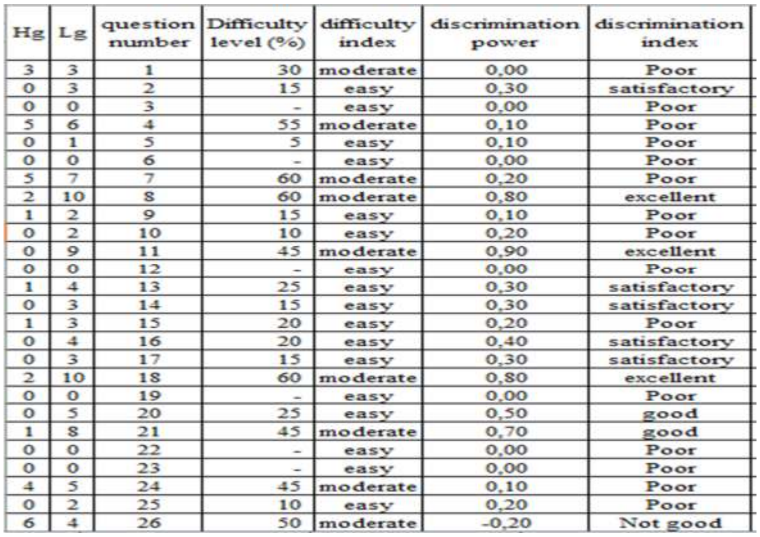
Table 2. High group and Low group



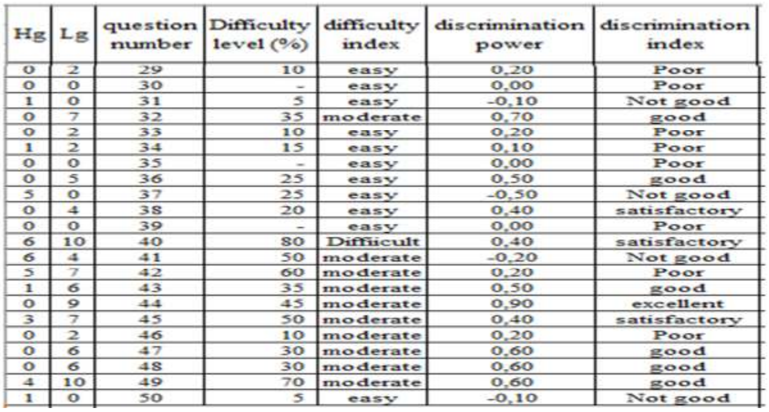
No		Details students' answers	Score	
			TRUE	FALSE
1	Highest group	DCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCCCDADA	49	1
2		BCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCBCDAAA	46	4
3		BCBCCDBDCDCBBDDCDBDACDCBDCCCDBCACADCBBCDBDBCBCDAAA	46	4
4		DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADDBCCCDADA	43	7
5		DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
6		DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
7		DCBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	42	8
8		ACBBCDADCDCBBDDCDBDACDCCDDDCDBCACADCABCADABCCCDADA	41	9
9		DCBCCDBBCDCBADDCDCDABDCADCCADBCACADCBBCAAACCCCDADA	40	10
10		DCBCCDBBADCBBDCCDCDACDCCDDCBDBCBCBDCBBCDBDBCBCDAAB	39	11
11	Lower group	CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
12		CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
13		CCBDCDCCCDDBBDCDCCDABDCCDCDADBCACADABBCABDAADCDBAA	31	19
14		DCBCCDBCCDABCDDADCCCCDCBBBCAABCDCADABBCABCBDCCBAAA	31	19
15		DCBDCDBBAADBCDDCDCDAADCCDDCADBCBBDDCBBCABBBBCDBBAA	30	20
16		DCBDCDBBAADBCDDCDCDAADCCDDCADBCBBDDCBBCABBBBCDBBAA	30	20
17		DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
16		DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
19		DABCCDABCDDBBADCDCDBBDCADCCADBCCCADCBCCAACCABCBA A	30	20
20		DCBAADAACDCBADDCDCDBCDCBBACAABCDCADABCCBABBCBCDC A	29	21

Table 2 shows the group of high and low performers. Out of 36 students who took the test, 10 students were in high group and 10 students were in low group. The third step was recapitulation of difficulty level and discrimination power. After deciding the high and low performers, the recapitulation of difficulty level and discrimination power was done by analyzing total students who answered the question incorrectl. The result of the analysis are presented in the following table.

Table 3. Recapitulation of the difficulty level and discrimination power



Hg	Lg	question number	Difficulty level (%)	difficulty index	discrimination power	discrimination index
3	3	1	30	moderate	0,00	Poor
0	3	2	15	easy	0,30	satisfactory
0	0	3	-	easy	0,00	Poor
5	6	4	55	moderate	0,10	Poor
0	1	5	5	easy	0,10	Poor
0	0	6	-	easy	0,00	Poor
5	7	7	60	moderate	0,20	Poor
2	10	8	60	moderate	0,80	excellent
1	2	9	15	easy	0,10	Poor
0	2	10	10	easy	0,20	Poor
0	9	11	45	moderate	0,90	excellent
0	0	12	-	easy	0,00	Poor
1	4	13	25	easy	0,30	satisfactory
0	3	14	15	easy	0,30	satisfactory
1	3	15	20	easy	0,20	Poor
0	4	16	20	easy	0,40	satisfactory
0	3	17	15	easy	0,30	satisfactory
2	10	18	60	moderate	0,80	excellent
0	0	19	-	easy	0,00	Poor
0	5	20	25	easy	0,50	good
1	8	21	45	moderate	0,70	good
0	0	22	-	easy	0,00	Poor
0	0	23	-	easy	0,00	Poor
4	5	24	45	moderate	0,10	Poor
0	2	25	10	easy	0,20	Poor
6	4	26	50	moderate	-0,20	Not good

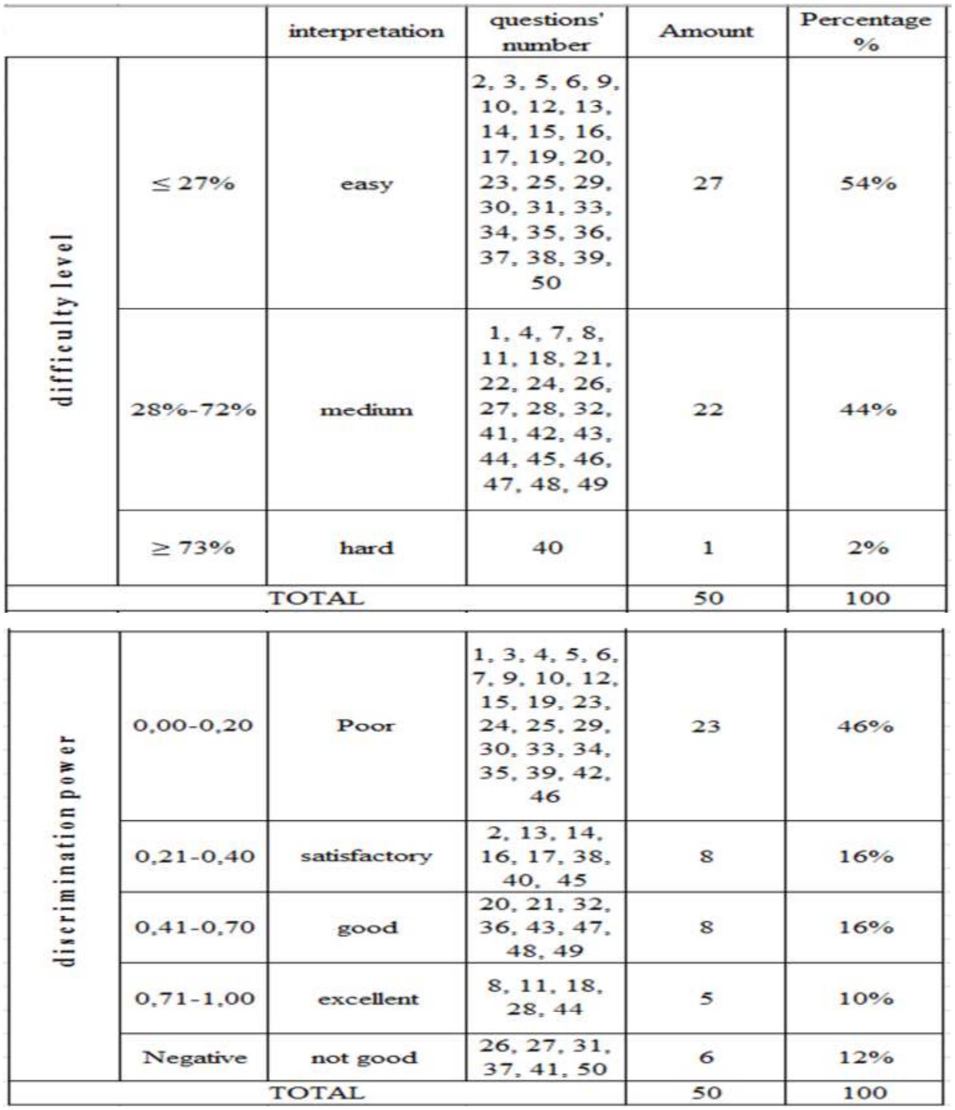


Hg	Lg	question number	Difficulty level (%)	difficulty index	discrimination power	discrimination index
0	2	29	10	easy	0,20	Poor
0	0	30	-	easy	0,00	Poor
1	0	31	5	easy	-0,10	Not good
0	7	32	35	moderate	0,70	good
0	2	33	10	easy	0,20	Poor
1	2	34	15	easy	0,10	Poor
0	0	35	-	easy	0,00	Poor
0	5	36	25	easy	0,50	good
5	0	37	25	easy	-0,50	Not good
0	4	38	20	easy	0,40	satisfactory
0	0	39	-	easy	0,00	Poor
6	10	40	80	Diffiicult	0,40	satisfactory
6	4	41	50	moderate	-0,20	Not good
5	7	42	60	moderate	0,20	Poor
1	6	43	35	moderate	0,50	good
0	9	44	45	moderate	0,90	excellent
3	7	45	50	moderate	0,40	satisfactory
0	2	46	10	moderate	0,20	Poor
0	6	47	30	moderate	0,60	good
0	6	48	30	moderate	0,60	good
4	10	49	70	moderate	0,60	good
1	0	50	5	easy	-0,10	Not good

Table 3 indicates recapitulation of the difficulty level and discrimination power. In terms of difficulty level, it was revealed 1 question was difficult, 21 questions were moderate and 28 questions were easy. In terms of discrimination power, 5 questions were excellent, 8 questions were good, 8 questions were satisfactory, 23 questions were poor, were 6 negative or not good. Questions number 26, 27, 31, 37, 41, 50 got negative discrimination index, which means that these questions were difficult for the high group but easy for the lower group. In other words, the lower group answered the questions correctly, while low groups did not. Question number 1, 3, 6, 12, 19, 22, 23, 30, 35, and 39 got 0.00 discrimination index which means that both groups can answer the questions correctly so that the questions were not acceptable because they could discriminate high performers from low performers. Meanwhile, questions number 8, 11, 18, 28, 44 were difficult for the low group but easy for the high group. So, 16 questions were poorly constructed because they got minus (-) index and 0.00 index. Therefore, those questions were unacceptable or have to be revised.

The fourth step was interpretation of difficulty level and discrimination power. In terms of difficulty level, the questions were analyzed as easy, moderate, and difficult items. While for discrimination power, the questions were analyzed as poor, good, satisfactory, excellent and not good (negative) items. The table below presents the summary of these categories.

Table 4. Interpretation of difficulty level and discrimination power



		interpretation	questions' number	Amount	Percentage %
difficulty level	≤ 27%	easy	2, 3, 5, 6, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20, 23, 25, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 50	27	54%
	28%-72%	medium	1, 4, 7, 8, 11, 18, 21, 22, 24, 26, 27, 28, 32, 41, 42, 43, 44, 45, 46, 47, 48, 49	22	44%
	≥ 73%	hard	40	1	2%
TOTAL				50	100
discrimination power	0,00-0,20	Poor	1, 3, 4, 5, 6, 7, 9, 10, 12, 15, 19, 23, 24, 25, 29, 30, 33, 34, 35, 39, 42, 46	23	46%
	0,21-0,40	satisfactory	2, 13, 14, 16, 17, 38, 40, 45	8	16%
	0,41-0,70	good	20, 21, 32, 36, 43, 47, 48, 49	8	16%
	0,71-1,00	excellent	8, 11, 18, 28, 44	5	10%
	Negative	not good	26, 27, 31, 37, 41, 50	6	12%
TOTAL				50	100

Table 4 shows the interpretation of difficulty level and discrimination power. The table indicates that in terms of discrimination level, there were 27 easy questions (54%), 22 moderate questions (44%), and 1 difficult question (2%). In terms of discrimination power, 5 questions were excellent (10%), 8 questions were good (16%), 8 questions were satisfactory (16%), 23 questions were poor (46%), 6 questions were negative or not good (12%).

The diagram below shows the percentage of difficulty levels of the test items. According to Arifin (2009: 270), a well-constructed should be not too easy or too difficult. To obtain good learning achievement, the proportion between the difficulty levels of the questions is spread evenly as the following options:

1. 25% of the items are difficult, 50% of the items are moderate, 25% of the items are easy.
2. 20% of the items are difficult, 60% of the items are moderate, 20% of the items are easy.
3. 15% of the items are difficult, 70% of the items are moderate, 15% of the items are easy.

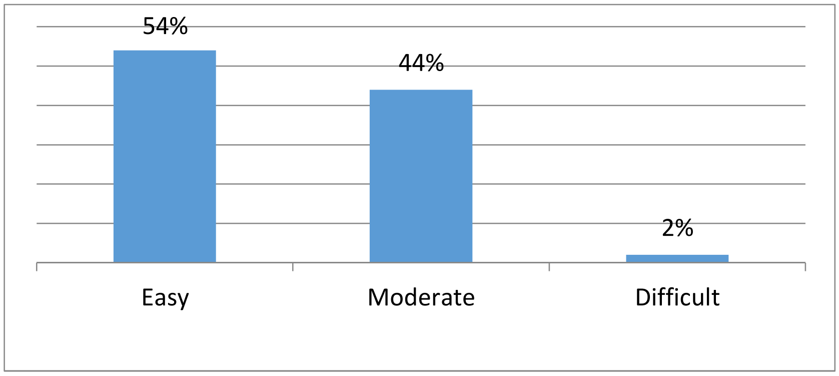


Figure 1. Percentage of Item Difficulty Level Diagram

The diagram (figure.1) above revealed that the English summative test was not well-constructed because 27 questions (54%) were easy, 22 questions (44%) were moderate and 1 question (2%) was difficult. Therefore, the English summative test did not meet the criteria of a good test. The diagram below shows the discrimination power of the test items. The discrimination power was calculated and interpreted (Arifin 2012: 270) as follows: the index of negative or not good item was minus, the index of poor items was between 0.00 and 0.20, the index of satisfactory items was between 0.20 and 0.40, the index of good items was between 0.40 and 0.70, the index of excellent item was between 70 and 1.00

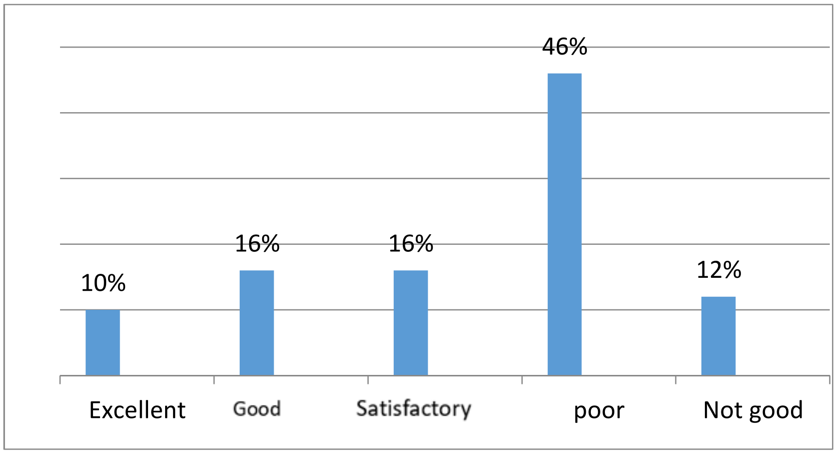
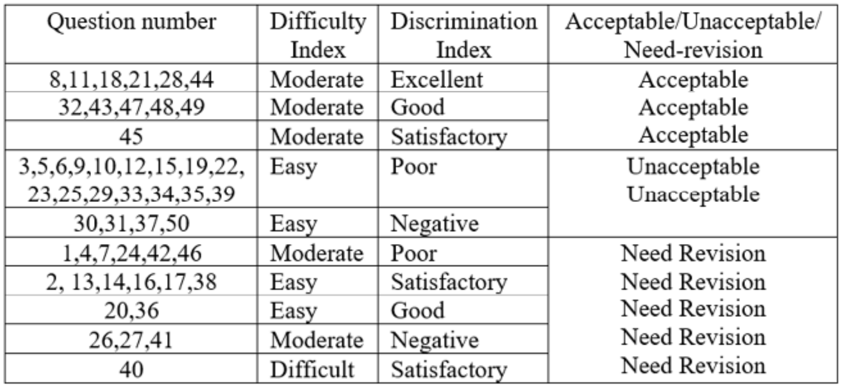


Figure 2. The percentage of item discrimination power

The diagram (figure. 2) above shows that 5 items (10%) were excellent because the discrimination index is between 70 and 1.00, 8 items (16%) were good because the index is between 0.40 and 0.70, 8 items (16%) were satisfactory because the index is between 0.20 and 0.40, 23 items (46%) were poor because the index is between 0.00 and 0.20, and 6 questions (12%) were negative or not good because the index was minus.

The last step was analyzing which items were acceptable or well-constructed and which items were unacceptable or ill-constructed. This final step is important for the recommendation because it will help teachers to revise the test. According to Arifin (2012) there were criteria to decide whether the test items are acceptable, unacceptable or need revision. The item is acceptable when the difficulty index is moderate and the discrimination index is ranging from satisfactory to excellent. The item is unacceptable when the difficulty index is easy or difficult and the discrimination index is poor or negative. While, the item has to be revised when the difficulty index and discrimination index vary where one of the indexes is the lowest in the category.

Table 5. The analysis of acceptable, unacceptable or need-revision items



Question number	Difficulty Index	Discrimination Index	Acceptable/Unacceptable/ Need-revision
8,11,18,21,28,44	Moderate	Excellent	Acceptable
32,43,47,48,49	Moderate	Good	Acceptable
45	Moderate	Satisfactory	Acceptable
3,5,6,9,10,12,15,19,22, 23,25,29,33,34,35,39	Easy	Poor	Unacceptable
30,31,37,50	Easy	Negative	Unacceptable
1,4,7,24,42,46	Moderate	Poor	Need Revision
2, 13,14,16,17,38	Easy	Satisfactory	Need Revision
20,36	Easy	Good	Need Revision
26,27,41	Moderate	Negative	Need Revision
40	Difficult	Satisfactory	Need Revision

Table 5 indicates the analysis of acceptable, unacceptable or need-revision items. It was revealed that 12 items were acceptable, 20 items were unacceptable and 18 items needed revision. The acceptable items must be used for the test, while unacceptable items should be dropped and new items should be constructed to replace them. The items that need revision should be recheck for further improvement. The revision could be done by rechecking the key answers, stems, distractors, teaching materials.

4. CONCLUSION AND SUGGESTION

Multiple choice is one of the most objective tests. In addition, it is less time consuming and easy to administer (Higgins & Tatham, 2003). Item analysis is essential in developing a test because it shows which items should be included, improved or eliminated (Gajjar, Kumar, & Rana, 2014). Based on data analysis of this study, it was found that 54% of test items (22) were easy, 44% of test items (27) were moderate, 2% of test items (1) was difficult questions. This means that most of the test items were easy and moderate in which both high stake and low stake students could answer them correctly. For the discrimination power, it as was revealed that 5 items (10%) were excellent, 8 items (16%) were good, 8 items (16%) were satisfactory, 23 items (46%) were poor, and 6 items (12%) were negative. Therefore, 12 items were acceptable, 20 items were eliminated and 18 items needed revisions. To sum up, this English summative test constructed by a junior high school teacher did not meet the criteria of a good test because it had poor item difficulty index and low discrimination index.

For further implication, the analysis of test items is necessary for evaluating the education system (Arhin, 2017). Beside eliminating misleading test items, it is also used to evaluate the quality of educational system (Malau-Aduli & Zimitat, 2011). If the tests are poorly constructed, this reflects low assessment literacy of the teacher. This study provides valuable insight for further item modification and test development. Quality control is important to make sure that the teachers produce good tests. They can seek advice from experts to validify their tests.

Based on the findings, some recommendations are made for future development of test items. First, the test item should be constructed by considering the criterion of a good test. Second, teachers need to seek experts to validate their test before administering it to the students. Third, the schools and government need to provide assessment literacy training for the teachers so that they have knowledge on making valid, reliable and practical test items.

REFERENCES

Arhin, A. K. (2017). Using Reliability and Item Analysis to Evaluate a Teacher-Developed Test in Educational measurement and Evaluation. Cogent Education, 4(1).

Arifin, Z. (2012). *Penelitian Pendidikan: Metode dan Paradigma Baru*. Bandung: Remaja Rosda Karya.

Arikunto, Suharsimi. (2008). *Penelitian Tindakan Kelas.* Bandung : Bumi Aksara.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Baker, B. F. (2001). *The Basics of Item Response Theory*. New York: ERIC.

Bodner, G. M. (1980). *Statistical Analysis of Multiple-Choice Exams*. Journal of Chemical Education, 188 - 190.

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

Djamarah. (2005). *Strategi Belajar Mengajar.* Jakarta: Rineka Cipta

Danuwijaya, A. A. (2018). *Item Analysis of Reading Comprehension Test for Post-Graduate Students*. English Review: Journal of English Education, 29 - 40.

Ding, L., & Beichner, R. (2009). *Approaches to data analysis of multiple-choice questions*. Physical Review Special Topics-Physics Education Research, 5(2), 1 - 17.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. New York: Routledge.

Gajjar, S., Kumar, P., & Rana, M. (2014). *Item and test analysis to identify quality multiple choice questions from an assessment of medical students of Ahmedabad, Gujarat*. Indian Journal of Community Medicine, 39(1), 1 - 17.

Henning, G. A. (1987). *A guide to language testing: development, evaluation research*. London: Newbury House Publisher.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum Associates.

Higgins, E., & Tatham, L. (2003). *Exploring the potentials of multiple-choice questions in assessment*. Learning and Teaching in Action, 2(1), 1 - 12.

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

Kehoe, J. (19951). *Basic item analysis for multiple choice tests. Practical Assessment, Research & Evaluation,* 4(10), 1 - 10.

Malau-Aduli, B. S., & Zimitat, C. (2011). *Peer review improves the quality of MCQ examinations. Assessment & Evaluation in Higher Education*, 37(8), 919 - 931.

McNamara, T. (2000). *Language Testing. Oxford*: Oxford University Press.

Plake, B. S., & Impara, J. C. (1996). *Handbook of Classroom Assessment*. In G. D. Phye. London: Elsevier.

Sudijono. (2009). *Pengantar evaluasi pendidikan*. Jakarta: Raja grafindo.

Toksoz, S., & Ertunc, A. (2017). *Item Analysis of a Multiple-Choice Exam. Advances in Language and Literary Studies*, 141 - 146.

Woodford, P. E. (1980). *Foreign Language Testing*. The Modern Language Journal, 64(1), 97 - 102.