EduLite *Journal of English Education, Literature, and Culture*
Vol. 6, No. 2, August 2021, pp. 256-269

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test

[1]Sayit Abdul Karim*, [1]Suryo Sudiro, [2]Syarifah Sakinah
[1]English Education Department, Universitas Teknologi Yogyakarta, DIY, Indonesia
[2]SMP N 2 Kempo, West Nusa Tenggara Province, Indonesia

***Corresponding Author**
Email: sayit.a.k@uty.ac.id

Received:	Revised:	Accepted:	Published:
30 March 2021	30 July 2021	1 August 2021	31 August 2021

Abstract

Apart from teaching, English language teachers need to assess their students by giving a test to know the students' achievements. In general, teachers are barely conducting item analysis on their tests. As a result, they have no idea about the quality of their test distributed to the students. The present study attempts to figure out the levels of difficulty (LD) and the discriminating power (DP) of the multiple-choice (MC) test item constructed by an English teacher in the reading comprehension test utilizing test item analysis. This study employs a qualitative approach. For this purpose, a test of 50-MC test items of reading comprehension was obtained from the students' test results. Thirty-five students of grade eight took part in the MC test try-out. They are both male (15) and female (20) students of junior high school 2 Kempo, in West Nusa Tenggara Province. The findings revealed that16 items out of 50 test items were rejected due to the poor and worst quality level of difficulty and discriminating index. Meanwhile, 12 items need to be reviewed due to their mediocre quality, and 11 items are claimed to have good quality items. Besides, 11 items out of 50 test items were considered as the excellent quality as their DP scores reached around 0.44 through 0.78. The implications of the present study will shed light on the quality of teacher-made test items, especially for the MC test.

Keywords*: discriminating power; item analysis; level of difficulty; reading comprehension test; teacher-made test*

INTRODUCTION

Language assessment holds a pivotal role in the field of education. Practically, teachers must assess their students at the end of learning and teaching process to know their students' learning progress and learning outcomes (Luthfiyyah et al., 2021; Hartati & Yogi, 2019). Generally, in the Indonesian context, teachers use tests to assess students at the end of the learning process what a so-called summative test (Maharani & Putro, 2020). On this ground, assessment is needed to find out students' achievement in a given

domain by using an instrument which is called a 'test'. The result of a test provides stakeholders with education on various aspects of assessment. By using a well-constructed item, teachers can assess students' competencies in the given domain.

Conceptually, different tests serve different objectives and it is believed that no single test may be developed and used to serve many different purposes (Sulistyo, 2018). There are some types of tests; based on their method, purpose, and the nature of the answer which can be given to students. One of them is the multiple-choice (MC) in the reading comprehension tests. According to Hemmati and Ghaderi (2014); Jayanti et al., 2019), MC is the most well-known testing item in the side classroom. The quality of MC is determined by the level of its validity, reliability, and discrimination ability (Jannah et al., 2021; Manalu et al., 2019; Arikunto, 2013). MC test type has been worldwide used in all over the world, including in Indonesia. Meanwhile, Mahmud (2014) affirms that most language examinations these days have MC tests, and it attracts the attention of many scholars in different countries and contexts.

Our intensive discussion with an English teacher, who made the MC test to assess the students' learning outcomes in the junior high school 2 Kempo, revealed that she has never conducted the item analysis for her own-made English MC test in the reading comprehension test. Therefore, she has no idea about the quality of the test she has constructed. To this end, it is important to examine the index level of difficulty (LD) and level of discrimination power (DP) of the MC test item constructed by the teacher in the reading comprehension test to figure out the quality of the test items. Besides, the findings of item analysis will be used as a shred of empirical evidence which in turn helps the teacher develop a good and useful item bank for practical utility. In short, by doing the test item analysis we can figure out that test the items are effectively evaluating students' learning progress.

The present study is necessary to be conducted because the findings provide some benefits to the teachers and related parties, concerning the MC test items construction and test item quality. The item analysis helps teachers and test developers to evaluate students' learning competence which in turn interprets students' progress at the end of the learning process. Moreover, test item analysis provides high-quality test items especially for a reading comprehension MC test because it increases the validity and reliability of the test. Besides, item analysis enables teachers to decide whether or not to use the items; revise or drop the poor items so that the teachers possess very high-quality test items before distributing them to students in the reading comprehension test.

The present study attempts to examine the LD and DP of MC items constructed by an English teacher in the English reading comprehension test for junior high school 2 Kempo, in West Nusa Tenggara Province. It is worth conducting this study because the findings will be used by the English language teachers when conducting reading comprehension tests using MC test type, especially for the junior high school students in the aforementioned institution.

EduLite *Journal of English Education, Literature, and Culture*
Vol. 6, No. 2, August 2021, pp. 256-269

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

For teachers, as test developers, the item analysis is worth conducting and may take advantage of accomplishing item analysis. This is in line with Sulistyo (2018), who states that the advantages may concern the quality of the items the teachers develop, and through item analysis, the item level will be identified as easy, moderate, or difficult. Item analysis is necessary to be conducted in order to have a high-quality test item that will be used in the next assessment period (Quaigrain & Arhin, 2017). According to Manalu et al. (2019), item analysis can function as information about items that should be improved or eliminated due to their poor quality. Meanwhile, Ingale et al. (2017) states that item analysis is the process of gathering information from students' responses to know the quality of test items. According to Sulistyo (2018), item analysis is a necessary stage in the test construction. It plays a vital role to examine the quality of individual items in a test, in particular of the multiple-choice type. Item analysis can help us analyze the effectiveness of our test questions.

The aim of conducting item test analysis is to examine the poor items, difficulty level, and validity and reliability of the test (Fitrianawati, 2010; Suprananto, 2012). In general, item analysis is used to figure out whether or not the items have good quality test items. According to Boopathiraj and Chellamani (2013), the item analysis of a test usually comes after the test developers construct, administer, and scored them out.

Teachers often assume that the test item which they have made is already good. As a result, if their students' scores are not satisfactory, they just take a conclusion that it is their students who have not mastered the materials well, while the success of one test is determined not only by the students but also on the quality of the test itself. According to Hartoyo (2011), teachers often assume that the test item which they have made are already good. As a result, if their students' scores are not satisfactory, they just take the conclusion that it is their students who have not mastered the materials well, while the success of one test is determined not only by the students but also on the quality of the test itself.

That is why teachers need to try out the already made test items and they must evaluate the quality of their test items and analyze test results as well as their construction. Then it is expected that the result of the analysis is carefully reviewed or revised and finally they will be used for the real test items. According to Bacon (2003), test score will be reliable through a well-constructed test and the test will cover a wide range of topics in the course. Thoughtfully written MC items can serve to evaluate students' cognitive processes (Buckles & Siegfried, 2006; Palmer & Devitt, 2007).

The analysis of multiple choices would provide information about the index of difficulty of test items, the level of discrimination of test items, and the effectiveness of each option. Gronlund (1993), suggested the item analysis procedures:

a) Arrange all (number of test papers) in order from the highest score to the lowest score.
b) Select about 1/3 of papers with the highest scores and take this as an upper group (number papers). Select the same number of papers with the lowest scores and take this as a lower group (number papers). Set the middle group of papers aside (the rest of the papers). Although

these could be included in the analysis, using only the upper and lower groups simplifies the procedure.

The other way in separating the groups is to classify the result into 27% of students with the highest test for the upper group and 27% of students with the lowest score for the lower group.

c) For each item count the number of students in the upper group who selected each alternative to teaching count refers.
d) Record the count from step 3 on a copy of the test, in columns to the left of the alternatives to which count refers.

Level of Difficulty (LD)

According to Haladyna (2004), the purpose of conducting item difficult is to identify the percentage of students who answer correctly. The estimation of item difficulty can be done by determining the percentage of students who selected the correct response or did the correct item. The simplest procedures are to base this estimate only on those students included in the item-analysis groups. Thus sum the number of students in the upper group (Ru) and lower group (Rl); sum the number of students who selected the correct answer, and divide the first sum into the second and multiply by 100.

The following is the formula for computing item difficulty (P-value) as suggested by Gronlund (1993).

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

P: the percentage of test-takers who did the correct items.
R: the number of test-takers who did the correct items.
T: total number of test-takers who tried the items.

The interpretation of the level of difficulty of the items is summarized in Table 1 as follows:

Table 1. The Category of Level of Difficulty (LD) of the Test Items

Index Range	Category	Frequency	Item Number
0-00.0.30	Difficult		
0.31-0.70	Moderate		
0.71-1.00	Easy		

Discriminating Power (DP)

The estimation of item DP can be done by comparing the number of students in the upper group (Ru) and the lower group (Rl who did the correct items. The discrimination power indicates how well the question separates the students who know the material well from those who do not. The analysis of the index of discrimination power may be counted by subtracting the amount of the lower group who did the correct items from the amount of the upper group who did the correct items and divided by the number in each group.

The following is the formula for computing discriminating power as suggested by Gronlund (1993).

EduLite *Journal of English Education, Literature, and Culture*
Vol. 6, No. 2, August 2021, pp. 256-269

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

$$D: \frac{Ru - Rl}{1/2\ T} \times 100$$

D: discriminating power index
Ru: the number in the upper group who did the correct items (among the 27% of those with the highest scores).
Rl: the number in the lower group who did the correct items (among the 27% of those with the lowest test scores).
½T: one-half of the total number of students included in the item analysis.

The DP of an item is presented as a decimal fraction: maximum positive DP indicated by an index of 1.00. This is obtained only when all students in the upper group answer correctly and no one in the lower group does. According to Backhoff et al. (2000), the rule of thumb for determining the quality of the items is the discriminating index. The values of D and their corresponding interpretation are shown in Table 2 as follows:

Table 2. Discriminating power of the answers according to their D value

D =	Quality	Recommendation
> 0.39	Excellent	Retain
0.30 – 0.39	Good	Possibilities for improvement
0.20 – 0.29	Mediocre	Need to check/review
0.00 – 0.20	Poor	Discard or review in-depth
< -0.01	Worst	Definitely discard

Adopted from Backhoff et al. (2000)

Multiple Choice Test Items

MC test is one of the test types that test takers are asked to select the best answer out of the choices from a list. Normally, the MC item has a stem and a set of options. Furthermore, the position of a stem is normally in the initial part of the item. The options are the possible answers that the test takers can select from, with the correct answer called the *key* and the incorrect answers called *distracters*. Moreover, the test takers need to select one best answer to the question provided. Indeed, item analysis can save much time and energy for both teachers and test developers. For these reasons item analysis is widely used to improve test item quality.

Several studies were carried out to examine the test items quality by using item analysis in the secondary settings. For instance, Hartati & Yogi (2019) conducted a study on item analysis for a better-quality test. They examined the document of teachers' English summative tests and students' answer sheets of SMA Muhammadyah Pamulang, Banten. The results of the study showed that the summative test has more easy items than difficult items. They went on to say that the proportion of the easy items is higher than expected. Besides, the discriminating power of some items is very poor. Another study was conducted by Manalu et al. (2019). They would like to figure out the quality of the reading final examination in SMA N 8 Medan which has been used by the students of grade nine. The findings revealed that more than half of the items constructed are valid and reliable test items.

In the junior high school context, a study on item analysis was carried out by Maharani and Putro (2020). They tried to analyze the English final semester test for junior high school students of grade nine in Ponorogo, East

Java. Their study revealed that the test does not have a good proportion among difficult, medium, and easy items. However, the findings gave insights that item analysis is a necessary process in constructing tests. The implication of the study is that teachers and test developers may take the benefit from the empirical evidence that quality test items received from well-constructed items through item analysis. A more recent study was conducted by Jannah et al. (2021). They examined multiple-choice items of English try-out using item analysis. It is an official exam paper they used as a document which they took from junior high school students' test results. The results of the study showed that the LD on the test items is varied. It was found out that some items were easy, moderate, and difficult to answer.

The previous study (Hartati & Yogi, 2019; Manalu et al., 2019; Maharani & Putro, 2020), yielded different results. Hartati & Yogi (2019) for instance, stated that the summative test has more easy items than difficult items. Besides, the proportion of the easy items is higher than expected. Meanwhile, in Manalu et al., 2019; Maharani & Putro's findings (2020), that the test does not have a good proportion among difficult, medium, and easy items. Those aforementioned previous studies portrait the importance of having a good quality of item analysis in MC item tests constructed by teachers, lecturers in the junior high school and senior high school settings.

Those previous studies have been caried out on the item analysis of MC test in the reading comprehension test both in the junior high schools and senior high school settings. However, very few of those studies, to the best of the authors' knowledge have a focus on examining the LD and DP in a teacher-made test utilizing test items analysis. Besides, the study on the importance of having a good quality test item of the teacher-made test is still very limited in numbers. On this account, the present study attempts to examine the LD and the DP of the MC test item constructed by an English teacher in the reading comprehension test utilizing test item analysis.

METHOD

This study employs a qualitative approach. For this purpose, a set of 50-multiple choice items tests of the reading comprehension was analysed based on the students' test results used as the source of data. Thirty-five students of grade eight took part in the try-out of the reading comprehension MC test. They are both male (15) and female (20) students of junior high school 2 Kempo, in West Nusa Tenggara Province. Moreover, these students know the basics of how to read and write and are expected to be able to read passages. It is ensured that students do not possess any reading disabilities. Besides, the answer sheets were also provided together with the reading comprehension test.

For the present study, there are thirteen simple reading texts developed and distributed to all respondents. The MC questions were developed by following the procedures of establishing a good MC test item construction such as; consist of two basic parts, namely a problem (stem) and a list of suggested solutions (alternatives), a number of incorrect or inferior alternatives (distractors). Furthermore, some steps were taken in developing the MC test items in order to have appropriate MC test item for junior high school students, they are; using familiar language, making sure there is only one best

EduLite *Journal of English Education, Literature, and Culture*
Vol. 6, No. 2, August 2021, pp. 256-269

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

answer, making the distractors appealing and plausible, making the choices grammatical consistent with the stem, and place the choices in some meaningful order.

The test materials are including several genres such as; report, narrative, procedures, explanation, news, letter and email, and announcement. The test items are relevant to their learning materials as they have already learned them. It took nineteen minutes for the respondents to complete the reading test and submitted right after the time is over. A descriptive analysis was used to analyze the data in the form of the MC test items. The MC test items were analyzed for their levels of difficulty (LD) and the discriminating power (DP).

The item analysis procedures used in this study are 27% of students with the highest score (upper group) and 27% of students with the lowest score (lower group) as suggested by Gronlund (1993: 103). The following steps were conducted after having the test results:

a) Provided tables for item analysis (excel sheet)
b) Computed the students' scores based on the number of correct answers
c) Computed the maximum, minimum, average scores, and put their ranks
d) Arranged the students' scores from the highest score to the lowest score. and rearrange their scores
e) Rearrange the students' scores into its ranks based on the top score to the lowest score
f) Selected 27% of the papers with the highest scores (upper group), and selected 27% of papers with the lowest scores (lower group). Thus, we can calculate the percentage; 27% x 35= 9,45, which means that we have 9 students for the upper group (Ru), and 9 students for the lower group (Rl).
g) Computed the maximum, minimum, and average scores
h) Inserted those figures into a table as formulated in table 3

RESULTS AND DISCUSSION

Results

To figure out the level of difficulty (LD) and the discriminating power (DP) of the reading comprehension test constructed by the English teacher, a try-out of 50 multiple choice test items was given to thirty-five grade eight students of junior high school 2 Kempo. The test items are relevant to their learning materials as they have already learned them. The test materials are including several genres such as; report, narrative, procedures, explanation, news, letter and email, and announcement. The data of students' test results of LD and DP can be summarized in Table 3.

Table 3. Level of Difficulty (LD) and Discriminating Power (DP)

ITEM	RU	RL	RU+RL	RU-RL	LD	DP	REMARK
					RU+RL / N	RU-RL / 0,5 N	
1	8	5	13	3	0,72	0,33	Good
2	7	5	12	2	0,67	0,22	Mediocre
3	5	1	6	4	0,33	0,44	Excellent
4	8	5	13	3	0,72	0,33	Good
5	6	3	9	3	0,50	0,33	Good
6	6	7	13	-1	0,72	-0,11	Worst
7	7	5	12	2	0,67	0,22	Mediocre
8	5	4	9	1	0,50	0,11	Poor
9	5	6	11	-1	0,61	-0,11	Worst
10	8	7	15	1	0,83	0,11	Poor
11	8	6	14	2	0,78	0,22	Mediocre
12	6	1	7	5	0,39	0,56	Excellent
13	7	5	12	2	0,67	0,22	Mediocre
14	9	6	15	3	0,83	0,33	Good
15	5	5	10	0	0,56	0,00	Poor
16	8	5	13	3	0,72	0,33	Good
17	8	5	13	3	0,72	0,33	Good
18	5	5	10	0	0,56	0,00	Poor
19	9	6	15	3	0,83	0,33	Good
20	6	4	10	2	0,56	0,22	Mediocre
21	8	5	13	3	0,72	0,33	Good
22	8	3	11	5	0,61	0,56	Excellent
23	7	3	10	4	0,56	0,44	Excellent
24	9	4	13	5	0,72	0,56	Excellent
25	8	6	14	2	0,78	0,22	Mediocre
26	8	7	15	1	0,83	0,11	Poor
27	6	5	11	1	0,61	0,11	Poor
28	9	4	13	5	0,72	0,56	Excellent
29	6	6	12	0	0,67	0,00	Poor
30	3	3	6	0	0,33	0,00	Poor
31	8	5	13	3	0,72	0,33	Good
32	9	6	15	3	0,83	0,33	Good
33	9	7	16	2	0,89	0,22	Mediocre
34	8	4	12	4	0,67	0,44	Excellent

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

35	2	3	5	-1	0,28	-0,11	Worst
36	7	6	13	1	0,72	0,11	Poor
37	9	7	16	2	0,89	0,22	Mediocre
38	3	3	6	0	0,33	0,00	Poor
39	6	8	14	-2	0,78	-0,22	Worst
40	8	6	14	2	0,78	0,22	Mediocre
41	9	6	15	3	0,83	0,33	Good
42	6	6	12	0	0,67	0,00	Poor
43	8	7	15	1	0,83	0,11	Poor
44	9	7	16	2	0,89	0,22	Mediocre
45	9	7	16	2	0,89	0,22	Mediocre
46	6	1	7	5	0,39	0,56	Excellent
47	9	2	11	7	0,61	0,78	Excellent
48	8	3	11	5	0,61	0,56	Excellent
49	4	2	6	2	0,33	0,22	Mediocre
50	9	4	13	5	0,72	0,56	Excellent

Table 3 provides information about the level of difficulty and discriminating power of the MC test items which was obtained from students' responses. As we can see those 16 items out of 50 test items were rejected due to poor and worst quality of the level of difficulty and discriminating index quality. The 16 rejected test items are consist of 4 worst quality and 12 poor quality test items. The worst quality test items can be seen in the test item numbers; 6, 9, 35, and 39, with the DP scores; -0,11, -0,11, -0,11, and -0,22 respectively. Meanwhile, the poor-quality test items can be seen in the test item numbers 8, 10, 15, 18, 26, 27, 29, 30, 36, 38, 42, and 43, with the DP scores; 0,11, 0,11, 0,00, 0,00, 0,11, 0,11, 0,00, 0,00, 0,11, 0,00, 0,00, and 0,11 respectively.

Furthermore, 12 items out of 50 test items need to be reviewed due to their mediocre quality. They are not very good items as the LD and DP scores are not met the standard of good items. The items that should be rechecked or reviewed are item numbers; 2, 7, 11, 13, 20, 25, 33, 37, 40, 44, 45, and 49, with all the same DP scores that are 0,22 respectively. Furthermore, the test items numbers 1, 4, 5, 14, 16, 17, 19, 21, 31, 32, and 41 are claimed to have good quality items as their LD and DP scores are met the quality of good test (0, 30-0,39). The DP scores in this try-out are 0,33 for all test items mentioned above. Those items were accepted, but they need to be improved and expected to reach their DP scores at least >0,39, so that they will be retained and used as the ready items to be distributed to the students.

The results of the test also showed that there were 11 items out of 50 test items considered as the excellent quality as their DP scores reached around 0, 44 through 0,78. Moreover, the items which are in the category of excellent can be seen in items 3, 12, 22, 23, 24, 28, 34, 47, 48, and 50. To sum up the test items quality, of 50 test items, 11 items are considered as excellent quality, 11 items are good, 12 items are in the mediocre level, 12 items are

poor, and 4 items are considered as the worst test items. The interpretation of the level of difficulty of the test items can be summarized in Table 4 as follows:

Table 4. The Category of Level of Difficulty (LD) of the Test Items

Index Range	Category	Frequency	Item Number
0-00.0.30	Difficult	1	35
0.31-0.70	Moderate	23	2, 3, 5, 7, 8, 9, 12, 13, 15, 18, 20, 22, 23, 27, 29, 30, 34, 38, 42, 46, 47, 48, 49
0.71-1.00	Easy	26	1, 4, 6, 10, 11,14, 16, 17, 19, 21, 24, 25, 26, 28, 31, 32, 33, 36, 37, 39, 40, 41, 43, 44, 45, 50

The interpretation of the discriminating power is summarized in Table 5 as follows:

Table 5. The Interpretation of Discriminating Power

Item Quality	Item Number	Frequency	Recommendation
Excellent	3, 12, 23, 24, 28, 34, 46, 47, 48, 50	11	Retain
Good	1, 4, 5, 14, 16, 17, 19, 21, 31, 32, 41	11	Possibilities for improvement
Mediocre	2, 7, 11, 13, 20, 25, 33, 37, 40, 44, 45, 49	12	Need to check/review
Poor	8, 10, 15, 18, 26, 27, 29, 30, 36, 38, 42, 43	12	Discard or review in-depth
Worst	6, 9, 35, 39	4	Definitely discard

Discussion

The present study tries to examine the level of difficulty (LD) and discriminating power (DP) of MC test items constructed by an English teacher (teacher-made test) in the English reading comprehension test for junior high school 2 Kempo, in West Nusa Tenggara Province. By conducting the test item analysis, the English teacher who constructed the items recognizes the quality of the MC test items she constructed. It is in line with Fitrianawati's (2010), who says that the objective of the test item analysis are to find out the bad item or items do not have good function, to increase item test through analysis of level of dificulty, and to increase the validity and reability of the teat item. Thus, a teacher could decide which items should be revised, dropped, and used for the final test items to assess the students' learning progress at the end of the teaching and learning session.

Test item analysis is very necessary to be conducted by test developers or teachers before distributing to students (Danuwijawa, 2018). It evaluates the performance of each function demonstrated in the test. Besides, it provides empirical evidence of the items that should be improved to achieve a good quality test item. Moreover, the test item analysis investigates the patterns of responses for both persons and items and describes the results of each test performed.

EduLite *Journal of English Education, Literature, and Culture*
Vol. 6, No. 2, August 2021, pp. 256-269

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

As stated previously, the reading comprehension test items are arranged based on the teaching and learning materials taught in grade eight of junior high school. The MC test items include many genres such as; procedures, reports, narrative, explanation, news, letter and email, and announcement since they have been learning the genres previously. The refore, the genres are choosen to achieve students' learning objectives. It is necessary for teachers to evaluate their students based on their learning objectives. Nisa & Helmanda (2020, p.82) confirm that in teaching and learning process, teachers should evaluate students by conducting a test and the test items must reach the students' learning objectives.

The findings of the present study revealed the quality of the MC test item constructed by the English teacher. Furthermore, the test items quality were reflected in the students' scores in the reading comprehension test. The results of the present study revealed that 16 items out of 50 test items were rejected since the difficulty level and discriminating power are in the category of poor and worst quality. Meanwhile, there are 12 items out of 50 test items that should be reviewed due to their mediocre quality. Moreover, there were 11 items out of 50 test items considered as the excellent quality since their DP scores reached about 0,44 through 0,78. The present finding does not support Manalu et al. s' (2019) finding showing that the number of multiple choices categorized as easy questions were 3 items (12%), satisfactory category 7 items (28%), difficult category 2 items (8%).

Nine (9) students from the upper group (Ru) and nine (9) students from the lower group (Rl) were picked up as the source of data for test analysis. The table shows that the upper group (9 students) were answered test items number 14, 19, 24, 28, 32, 33, 37, 41, 44, 45, 47, and 50 correctly. While none of the lower groups (9 students) answered them correctly. 8 students answered test item number 38 correctly, but they are very poor in question number 3, 12, and 46. Thus, test items number 14, 19, 24, 28, 32, 33, 37, 41, 44, 45, 47, and 50 distinguished between the upper group and the lower group on the test as a whole. Such items are called positive discriminators. The present finding is opposite with Hartati & Yogi's findings (2019) that the propostion of the easy items is higher than expected 19 (38%) and the level of discriminating power of the items are very poor. It means that the items could not distinguished between the lowe and the upper group (negative discriminators).

While test item numbers 15, 18, and 42 are considered as negative discriminators as they did not make any difference between the upper group and lower group (both groups have the same number of correct responses on the items). Meanwhile, Manalu et al.'s finding (2019) revealed that the distractors are able to distract more students in lower group.

The results of test analysis will be used for feedback on the test given and it provides precious information on how to organize a good test, and how to meet various types of educational targets. The results of test item analysis in the present study showed the tests items' capabilities and deficiencies for review and provide a means of assessing the next stage of development or testing. Besides, it provides a basis for assigning responsibility for deficiency correction.

We can identify the test items in terms of degrees of difficulty, namely easy, moderate, and difficult test items. Meanwhile, the discrimination power is in the form of worst, poor, mediocre, good, and excellent test items through the analysis. These test items quality information provide a quick overview of the test and can be used to identify items that are not performing well. Besides, such information can be used by teachers as the test developers to decide whether or not to retain, improve or discard the items. The test analysis provides an overview of question performance both in terms of the difficulty of questions and in terms of the discrimination of questions (upper and lower groups).

The tables of LD and DP (Table 3, 4, and 5) provide us very useful information about how the questions compose and assess students' performance. Moreover, in the table of item analysis, the numbers of correct responses from both groups tell us about the level of difficulty and also discrimination index respectively. The results of item analysis in the present study provide empirical evidence which items have good quality test items which are not in the reading comprehension MC test constructed by the English teacher for the students of junior high school 2 Kempo. It is in line with Hartati & Yogi's (2019, p.68), that the item analysis may help teachers as test developers to prepare high-quality test items in reading MC test in the future to ensure that the test items achieve their real objectives.

CONCLUSION

Conclusions

From the results of the analysis on the level of difficulty (LD) and discriminating power (DP), we can conclude that 16 items out of 50 test items were rejected due to the worst and poor-quality level. There were 12 items out of 50 test items that need to be reviewed due to their mediocre quality. They are not very good items as the LD and DP scores are not met the standard of good test items. The findings also showed that there were 11 items out of 50 test items considered as the excellent quality as their DP scores reached around 0,44 through 0,78. Besides, out of 50 test items, 12 test items must be reviewed since the items' quality is mediocre. The present study has some light on the quality of the teacher-made test and contributes to the body of knowledge on language assessment and testing.

Recommendations

The findings of this present study have significance for teachers, test constructors, and test developers in the related field. Therefore, it is highly recommended for teachers as the test developers and other stakeholders in the educational assessment to utilize the test item analysis to improve the existing teacher-made tests. Furthermore, the results of the item analysis may provide useful information about the quality of the test items constructed by English language teachers, so that the teachers and the related parties can judge and decide which test items must be revised, improved, and eliminated from the list before administrating them to the students. Finally, teachers and test constructors may replicate this item analysis in other subjects to develop high quality and useful item bank for practical utility.

E-ISSN: 2528-4479, P-ISSN: 2477-5304
http://jurnal.unissula.ac.id/index.php/edulite
DOI: http://dx.doi.org/10.30659/e.6.2.256-269

ACKNOWLEDGEMENTS

The authors would like to thank the students for their availability and willingness to involve in the data gathering process.

REFERENCES

Arikunto, S. (2013). *Dasar – dasar evaluasi pendidikan.* Bumi Aksara.

Backhoff. E.E, & Reyna, N.L, & Morales, M.R. (2000). The level of difficulty and discrimination power of the basic knowledge and skills examination (EXHCOBA). *Revista Electronica d Investigacion Educativa, 2,* (1), 1-16.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short asnwer question in a marketing context. *Journal of Marketing Education*, *25* (31-36). https://doi.org/10.1177/0273475302250570

Boopathiraj. C, & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & interdisciplinary Research, 2* (2).

Brown, D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education, Inc.

Buckles, S., & Siegfried, J.J. (2006). Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics. *Journal of Economic Education, 37* (48-57). https://doi.org/10.3200/jece.37.1.48-57

Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduatestudents. *English Review: Journal of English Education, 7*(1),29-40.https://journal.uniku.ac.id/index.php/ERJEE. https://doi.org/10.25134/erjee.v7i1.1493.

Fitrianawati, M. (2010). *Peran analisis butir soal guna meningkatkan kualitas butir soal, belajar peserta didik.* Seminar Nasional Pendidikan PGDS UMS & HDPGSDI Wilayah Jawa.

Gronlund, N.E. (1993). How to make achievement tests and assessments. University of Michigan.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items (3rd ed.). *Lawrence Erlbaum Associates Publisher.* https://doi.org/10.4324/9780203825945

Hartati, N., & Yogi, H.P.S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF), 2* (1), 59-70. Retrieve from: https://jurnal.umj.ac.id/index.php/ELIF. https://doi.org/10.24853/elif.2.1.59-70

Hartoyo. (2011). *Language assessment*. Pelita Insani

Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple- choice questions on the listening comprehension of EFL learners. *Procedia - Social and Behavioral Sciences*, *98*, 637–644. https://doi.org/10.1016/j.sbspro.2014.03.462.

Ingale, A.S, Giri, P.A., Mohan.K., Doibale. (2017) Study on item and test analysis of multiple choice questions amongst undergraduate medical students. *International Journal of Community Medicine and Public Health, 4 (5),1562-1565.* http://dx.doi.org/10.18203/2394-6040.ijcmph20171764.

Jannah, R., Hidayat, D.N., Husna, N., Khasbani, I. (2021). An Item analysis on multiple-choice questions: a case of a junior high Scholl English try-out test in Indonesia. *Leksika*, *15* (1), 9-17. https://dx.doi.org/10.30595/lks.v15i1.8768.

Jayanti, D., Husna, N., & Hidayat, D. N. (2019). The validity and reliability analysis of English national final examination for junior high school. *VELES Voices of English Language Education Society Journal*, *3*(2), 128–135. https://doi.org/10.29408/veles.v3i2.1551.

Luthfiyyah, R., Aisyah, & Sulistyo, G.H. (2021). Technology-enhanced formative assessment in higher education: A voice from Indonesian EFL teachers. *EduLite: Journal of English Education, Literature, and Culture*, *6* (1), 42-54. http://dx.doi.org/10.30659/e.6.1.42-54.

Maharani, A.V., & Putro, N.H.P.S. (2020). Item analysis of English final semester test. *Indonesian Journal of EFL and Linguistics, 5* (2), 491-504). https://doi.org/10.21462/ijefl.v5i2.302

Mahmud, M. (2014). The EFL students' problems in answering the Test of English as a Foreign Language (TOEFL): a study in Indonesian context. *Theory and Practice in Language Studies*, *4*(12), 2581-2587. https://doi.org/10.4304/tpls.4.12.2581-2587.

Manalu, D., Sipayung, K.T., Lestari, F.D. (2019). An analysis of students reading final examination by using item analysis program on eleventh grade of SMA Negeri 8 Medan. *Journal of English Language Teaching & Applied linguistics, 1* (1), 13-19. https://doi.org/10.36655/jetal.v1i1.98.

Nisa, R., & Helmanda, C.M. (2020). Analysis of reading comprehension final test at English Department of Muhammadyah Aceh University. 7 (1), 72-85. https://doi.org/10.46244/geej.v7i1.987

Palmer, E.J., & Devitt, P.G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? *BMC Medical Education, 7* (49). https://doi.org/10.1186/1472-6920-7-49

Quaigrain, K., & Arhin. A.K. (2017) Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation, *Cogent Education*, *4*,1-11. https://doi.org/10.1080/2331186X.2017.1301013.

Sulistyo, G. H. (2018). *EFL learning assessment at School*: An introduction to its basic concepts and principles. CV Bintang Sejahtera.

Suprananto, K. (2012). *Pengukuran dan penilaian pendidikan* [Measurement and assessment of education]. Graha Ilmu.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.