English Teaching Journal: A journal of English Literature, Linguistics, and Education
Vol.9, No.2, December 2021, pp. 143~152
ISSN: 2338-2678 DOI: 10.11591/etj.v9i2.10867

Utilizing Test Item Analysis to Portray the Quality of English Final Test

Nuri Ati Ningsih[1], Woro Widowati[1],
[1] Department of English Teaching, Universitas PGRI Madiun

Article Info

Article history:

Received November 3, 2021
Revised November 5, 2021
Accepted Aug 26, 2021

Keywords:

Item Analysis
English Item Test
Difficulty level
Discrimination index
Distractor

ABSTRACT

This study aims at describing the quality of the item test by analyzing the difficulty level, discriminating index, and distractor effectiveness of English final test item. The researchers use a descriptive quantitative as the research design. The population and sample of the study is the student's answer sheets of the tenth grade of SMK PGRI Wonoasri. There are 22 student's answer sheet as the primary data. The researchers use documentation technique and test kits of the English final test as the instrument. The data is analyzed by using classical theory measurement and item analysis' formula. The result shows: (1) the level of difficulty which categorized as difficult is 35% (40 items), medium is 47% (19 items), and easy is 18% (7 items); (2) discrimination index that recommended for being rejected is 15% (6 items), accepted 38% (15 items), and not accepted is 47% (19 items); (3) distractor effectiveness which has functioning distractors as very good is 20% (8 items), good is 43% (17 items), fair 27% (11 items), and less good is 10% (4 items). In short, the quality and acceptance of test items of English final is medium or fair. So, there are still a lot of items need to be revised.

Corresponding Author:
Nuri Ati Ningsih,
Departement of English Teaching, Universitas PGRI Madiun,
Jalan Setiabudi 85 Madiun, Jawa Timur, Indonesia.
Email: nuriatiningsih@unipma.ac.id

1. INTRODUCTION

In educational system, testing plays as an important role to convince whether the process of teaching and learning run properly or not. Testing refers to an effort done by the teacher to measure the result of the students in attending the teaching learning process. Test can be conducted along the process of teaching or at the end of teaching process. According to Brown (2004) test is a set of equipment to measure an individual's proficiency within particular criteria. The close definition of test proposed by Arikunto in 2016. She said that test is a type of procedure that can be used to determine and measure something in accordance to the way and the rules that have been set. It means that test is a tool that support the teachers to assess the students' competence in which the result can be used to predict the students' progress. Consequently, all the teachers should have ability to construct good test to get a result accurately.

Besides having well-constructed items, a good test has to meet a specific characteristics. Brown (2001) states that a good test should consist of at least three criteria. They involve practicality, validity and reliability. A test classified to be a good test if it meets some requirements, such as validity, reliability, practicability, objectivity and economical (Arikunto: 2016). Practicality means that the test should be prepared with the low budget, time limitation, implementation, and scoring system (Brown: 2001). Validity is the ability of the test in measuring what supposed to be measured dealing with the learning goals or competencies to be achieved. It refers to the extent to which an instrument really measures the objective to be measured and suitable with the criteria (Hatch and Farhady: 1982). On the other hand, reliability refers to the consistency of the test. This is to see the consistency of the test in measuring the test taker ability (Ali, et al: 2016). The test is claimed to be reliable when it taken by

the same person/students several times and there is no significant different in the result or test score achieved.

To ensure whether the test has good quality or not, the teachers or researchers must do item analysis. In analyzing test item, a good test at least should conform to three characteristics. Those are item difficulty, item discrimination, and effectiveness of distractors (Brown: 2004). Difficulty level is describing about how difficult or easy the item for the students. Brown (2004) writes that item difficulty relates to the number of the students who assume an item test difficult or easy. Detailed definition then proposed by Haladyna (2004) that item difficulty is conducted to identify the percentage of the students who can answer the test item correctly. According to Jannah et al (2021), the discrimination ability refers to the ability of the item test to set high – performing test takers apart from the lower-achieving counterpart. Discrimination power tells about the ability of the item test in discriminating the students who were classified in the upper and lower group. Distractor defines as the incorrect alternatives that test maker provides in the questions (Madsen: 1983). This characteristics can be analyzed on the form of multiple choice test types. An effective distractor must be chosen at least 5% of the test takers (Arikunto: 2016). Finding the effectiveness of distractor used to measure the functioning incorrect options in attract the students (Brown: 2004). Haladyna (2004) adds that analyzing distractor has several functions. It involves reducing item that use ineffective sentences or too many options, providing information to improve the items, assisting to choose a correct distractors, assisting to comprehend the students' cognitive behavior, and increasing the item response score.

Some previous studies have been done on discussing the quality of the item test by analyzing the level of difficulty, level of discriminant and finding the effectiveness of the distractors. Manulu (2019) determines the quality of the reading final examination. She used a mix method and analyzed the data by ANATES program. The result shows that many items were categorized as marginal and poor category in term of the level of difficulty, discrimination power and level of distractors. The similar study was conducted by Maharani et al in 2020. Their study sought to examine the quality of English Final Test. The data were analyzed based on item difficulty, item discrimination, and distractors' effectiveness by using Quest Program. The result shows that the English Final test does not have good proportion among easy, medium and difficult item. In the item discrimination, the test had excellent item and the distractors could distract the students because there 80 % of distractors were effective.

This study is addressed on describing the quality of the item test in the form of Multiple Choice item. The analysis represent on describing the difficulty level, discrimination power, and the effectiveness of the distractor of the test. The data of this study is analyzed by using classical theory measurement and item analysis' formula proposed by several experts. The result of this study is intended to offer a feedback for developing the quality of the test and also to give a real example for the teachers, test developers and students on how to analyze an item test to gain a good quality test.

2. RESEARCH METHOD

To describe the phenomenon within the teaching learning process, in this study, the researchers use descriptive quantitative. It this case, if we are going to represent individuals, groups, activities, events, or situations, then the most suitable research design to be used is descriptive quantitative (Leavy: 2017). Descriptive research is one of the type of quantitative research which is non-experimental. The quantitative research is required to use numbers. The use of numbers can be seen in the data collection, data interpretation or discussion and also in the result of the study.

The sampling technique used in this study is purposive sampling technique. The population of this study is the students' answer sheets of tenth grade of SMK PGRI Wonoasri in schooling year of 2019/2020 with the total amount is 90 sheets. It is taken from four classes which consist of 25 sheets of PM class, 22 sheets of MM class, 23 sheets of TKJ class, and 20 sheets of APK class. Within this study, the researcher uses sample by focusing on the student's answer sheets of MM class of the tenth grade of SMK PGRI Wonoasri which are consist of 22 sheets.

Through this study, the researchers uses documentation technique and uses documents as the main source data. The documents are the test kits of the English final test of the tenth grade of SMK

PGRI Wonoasri in the schooling year of 2019/2020. The data carried out in the form of primary data are in the form of copy hard file. The main instrument in this study are also the test kits of the English final test of the tenth grade of SMK PGRI Wonoasri. The test kits consist of key answer sheet, a question sheet, and answer sheets which have been taken from 22 students of class X MM.

English final test of the tenth grade of SMK PGRI Wonoasri was in the form of Multiple Choice Questions and Essays. The researchers only focused on the Multiple Choice Questions. The Multiple Choice Questions or the objective tests were analyzed by using the item analysis. The score for Multiple Choice Questions' answers would be 0 and 1. 0 was given for the wrong answers and 1 was given for the correct one.

3. RESULTS AND ANALYSIS

3.1. Difficulty Level

To determine the level of difficulty, the researchers accounted the item difficulty by using Brown formula. It can be done by calculating the proportion of the test takers who answer the test correctly.

$$DL = \frac{total\ number\ of\ students\ who\ answers\ correctly}{total\ numbers\ of\ test\ taker}$$

The scale of difficulty level ranges from 0.00 up to 1.00. Then, it is ranked in to several classifications. This classifications functions to interpret the level of the difficulties. The detail rank described in the following table:

Table 1. The Difficulty Level

Index level	Difficult Categories
0.00-0.30	Difficult
0.31-0.70	Moderate
0.71-1.00	Easy

(Brown & James: 1996) Table 2 shows the item difficulty index specifically. The detail distribution as the following:

Table 2. Classification of Item Difficulty

No	Difficulty Level	Test Item (Question Number)	Total (items)	Percentage	Category
1.	0,00 – 0,30	2, 6, 8, 9, 10, 14, 20, 22, 23, 25, 27, 35, 37.	14	35%	Difficult
2.	0,31 – 0,70	1, 3, 4, 5, 7, 11, 13, 18, 19, 24, 26, 29, 30, 31, 32, 33, 36, 38, 39, 40.	19	47%	Medium
3.	0,71 – 1,00	12, 15, 16, 17, 21, 28, 34.	7	18%	Easy

The table above indicates that the test item or item number 12, 15, 16, 17, 21, 28, and 34 are categorized as the *easy* one with the total number seven (7) items which is 18% of the total number of test items or multiple choice questions. At a meantime, test item number 1, 3, 4, 5, 7, 11, 13, 18, 19, 24, 26, 29, 30, 31, 32, 33, 36, 38, 39, and 40 are categorized as *medium* with 47% as the percentage from the total number of test items. The last one are item number 2, 6, 8, 9, 10, 14, 20, 22, 23, 25, 27, 35, and 37 which categorized as *difficult* items with 35% as the percentage from the total number of test items. As the results from these difficulty level data, the highest amount is in the level or categorization *medium* items with total amount nineteen items. The lowest amount is the level or categorization of *easy* items with total amount seven items. Overall, the highest percentage of difficulty level is within the *medium* category and the lowest percentage is within the *easy* category.

3.2. Descrimination Index

The level of discrimination index can be categorized into five different levels as the following table:

Table 3. The Discriminant Level

Index level	Difficult Categories
(negative)	Very poor
0.00-0.20	Poor Item
0.21-0.41	Satisfactory items
0.42-0.70	Good
0.71-1.00	Excellent

The results of those index is gained by measuring the level of discrimination by using Brown's formula and calculated manually.

$$DI = DL\ highest\ group - DL\ lowest\ group$$

The results of calculating the index of discriminant can be summarized through the table below:

Table 4. Distribution of Discrimination Index

No.	Discrimination Index	Test Item (Question Number)	Total (items)	Percentage	Category
1.	*Negative*	2, 5, 7, 13, 20, 25.	6	15%	Very Poor
2.	0,00 – 0,20	1, 4, 6, 8, 10, 11, 14, 15, 21, 22, 23, 24, 26, 27, 28, 31, 35, 39, 40.	19	48%	Poor
3.	0,21 – 0,40	3, 9, 12, 17, 32, 34, 36, 38.	8	20%	Satisfactory
4.	0,41 – 0,70	16, 18, 19, 29, 30, 33, 37.	7	18%	Good
5.	0,71 – 1,00	-	0	0%	Excellent

Table 4 above presents the data about a brief categorization and distribution of discrimination index on English final test. The first category with negative results means *very poor* items with the total number is six items number or 15% of the total amount of item number. Those items that included within this category are test item number 2, 5, 7, 13, 20, and 23. The second category which is *poor* items, has resulted nineteen items or 48% of the test items' total amount and categorized as the *poor* one. Those nineteen *poor* items, are test item number 1, 4, 6, 8, 10, 11, 14, 15, 21, 22, 23, 24, 26, 27, 28, 31, 35, 39, and 40. The third category is *satisfactory* items. The total amount in this category is eight items or 20% of the test items' total amount and they are test item number 3, 9, 12, 17, 32, 34, 36, and 38. Fourth category is *good* items with total amount that included in this category is seven items or 18% of the test items' total amount. Those items are test item number 16, 18, 19, 29, 30, 33, and 37. The last category is *excellent* items. Unfortunately, there is none of test items that belongs to this category.

From the description above, it can be concluded that the highest amount is nineteen items with 48% as the percentage from the total amount of the test items and it belongs to *poor* items categorization. The lowest amount that reveals is six items which belongs to the category *very poor* items with negative results. The percentage of this this distribution of discrimination index on English final test can be illustrated with the chart below:

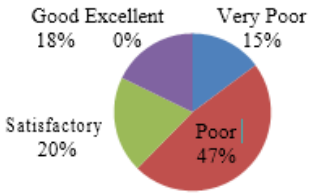

Figure 1.Percentage Distribution of Discrimination Index

3.3 Distractor Effectiveness

Distractor effectiveness can be seen from the distribution of students' answer patterns. The distribution of students' answer patterns describe on how the students choose the correct answer based on the given options. The options are consist of one correct answer and two, three, or four distractors. Distractors are functioning or effective if all of them are chosen by at least 5% of all the students (Arikunto: 2016). As a complete analysis results on distractor effectiveness, test items number 3, 9, 18, 20, 27, 35, 36, and 40 have all their distractors are functioning because each of the distractor is chosen by more than 5% of all the students. Meanwhile, test item number 1, 2, 7, 8, 10, 11, 12, 14, 19, 24, 25, 26, 29, 30, 33, 38, and 39 have three functioning distractors. Test item number 4, 5, 6, 13, 16, 21, 22, 23, 31, 32, and 37 have two functioning distractors. The rest of test item which are number 15, 17, 28, and 34 have only one functioning distractors. As a results, there are one hundred and nine (109) distractors which are functioning or 68% of the total 160 distractors on 40 test items. Whilrest of it is non-functioning distractors with total amount fifty- one (51) distractors or 51% of the total 160 distractors on 40 test items. These results can be summarized into the following chart:

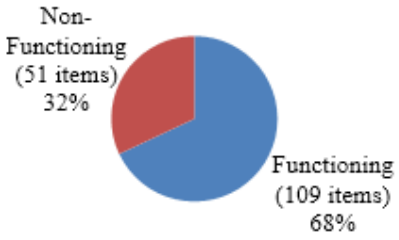

Figure 2. The Distribution of Distractor Effectiveness

3.4 Difficulty Level

Difficulty level is a proportion of how many students who give the correct answer from the total number of the students who take the test. A good test item is an item that is not too easy or too difficult. It can be said that a good test item is an item that has the difficulty level in *medium* categorization.

The result of this study shows that test items which categorized as *difficult*, have a total amount 14 items or 35% of the total test items. Test items which categorized as *medium* have a total amount 19 and 47% as the percentage of the total test items. The last categorization which is *easy* items, have 7 as the total amount and 18% as the percentage of the total test items. Therefore, it can be sum up that the test item on English final test for the tenth grade of SMK PGRI Wonoasri has not proportional item difficulty. This test is dominated by good item test that consist of 19 items or 47% of the total test items. Brown (2004) argued that a well-constructed item cannot be easy or difficult. It should cover each difficulty level so that teachers can recognize the abilities of each students.

These results are balanced to the previous item analysis studies on the level of difficulty of multiple choice item test. Manulu (2019) discovered that the package of the item test used in term of difficulty level is not good. There is no equal distribution on the range of difficulty level. The detail result show that the number of item categorized as easy amounted to 12 %, satisfactorily 38%, difficult 8 %, very difficult 12 % and very easy 32%. On the same matter, Jannah et al (2021) found that 14% of the item classified into easy, 66% and 20% categorized into moderately difficult and difficult item. In short, the level of difficulty of the item test are not evenly distributed.

Some factors can affect the index of difficulty of an item test. First, the direction or instruction given in a test convey big impact to the result. Different comprehension having by the students will influence their answers automatically. Second, the item test which are not relevant with the material given or discussed in the process of teaching also give impact on the index of difficulty. Finally, technique of examination also give contributions to the index value. The current situation in pandemic era, on line test tend to be applied than off line test. One of the weaknesses of on line test is in controlling the students' performance and the situation around them when joining the test. The teachers have low authority in running the test.

3.5 Descriminant Index

The function of measuring the discriminating index on multiple choice questions is to distinguish the students' abilities. The measurement in this stage requires the classification of the students into groups of students who take the test which can be obtained by calculating 27% of the highest and lowest scores. For the items which considered to be *accepted* is if the level of discrimination has an index ≥ 0,25. While, an item that has an index of discrimination less than 0,25 (<0,25) is considered for being not accepted and recommended for revision. In the meantime, negative discrimination index is supposed to be rejected (Kashap: 2015). The further classification of the results of discrimination index can be seen through the following table and chart:

Table 5. Recommended Results of Discrimination Index

No.	Discrimination Index	Test Item (Question Number)	Total (item)	Percentage	Note
1.	*Negative*	2, 5, 7, 13, 20, 25.	6	15%	Rejected
2.	≥ 0,25	3, 9, 12, 16, 17, 18, 19, 29, 30, 32, 33, 34, 36, 37, 38.	15	38%	Accepted
3.	< 0,25	1, 4, 6, 8, 10, 11, 14, 15, 21, 22, 23, 24, 26, 27, 28, 31, 35, 39, 40.	19	47%	Not Accepted, need to be revised
	Total		40	100%	

These results of the distribution above can be simplified into the following chart:

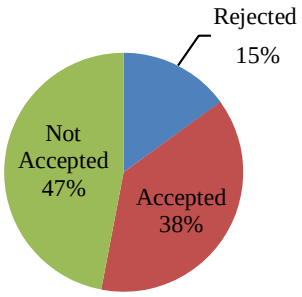

Figure 3. Recommended Results of Discrimination Index

Table 5 and figure 3 illustrate the data regarding to the recommendation of the results from discrimination index of English final test for the tenth grade of SMK PGRI Wonoasri. It can be seen that there are six items or 15% of the total test items which are rejected. Those items are test item number 2, 5, 7, 13, 20, and 25. It means that these item are supposed to be replaced with a new one.

In the meantime, there are 38% or fifteen items which are test items number 3, 9, 12, 16, 17, 18, 19, 29, 30, 32, 33, 34, 36, 37, and 38, that are all accepted. These test items can be stored in the bank of question so that it can be used in future. The last items which are 19 items or 47% that are not accepted and to be revised. These items do not need to be replaced with the new one. Overall, based on the results of discrimination index on English final test, the items on multiple choice questions are mostly less good and should be revised. Only few of them that need to be replaced because of the poor quality.

An item test which has less good value on discrimination power can be identified by observing the words put in an item test. Ambiguous wording used in item test can trigger a lower and negative value on discrimination power (the office of Educational Assessement: 2016). Ambiguous wording on test is very dangerous since it is not only can lead the students to choose an incorrect alternative or options but also avoid the students in having full understanding of what a question or an item expects them to do. Accordingly, a revision on an ambiguous wording used in item test should be done to eliminate the error or mistake stated in an item test and also to improve the quality of the item test.

The previous study dealing with the various level of discriminating power has been done. Jannah, et al (2001) reported the result of their study that there is no item test with excellent discriminating power. However, there are 16 and 9 items with good and satisfactory level of discriminant, 16 item with poor level of discriminant and 9 item should be rejected because of negative

scores. In Sum, half of the question or item test on trial testing should be removed or at least revised and improved.

3.6. Distractor Effectiveness

One of the characteristic of multiple choice questions is having one question and some options of answers or alternatives. Among those options, there is only one correct answer. While the false answers are called the distractors. Classifying the distractor in to effective or not, the test makers need specific standard. A test item is considered to be a good item when an item has three functioning distractors which each of the distractor is chosen by at least 5% of the students. While test item which categorized as fair item is an item that has two functioning distractors. An item which has only one functioning distractors is categorized as less good test item.

The last category is *not good* item which recognized when all of the distractors are not functioning (Arikunto: 2016). The similar prior research has been done dealing with an ineffective distractor. Hartati and Yogi (2019) reported their result of research that the distractor of an item test in their study arised to be totally unrelated to the question and the distractors failed to attract the test takers to choose them.

In this research, there are 40 test items with the total amount of distractors are 160. The obtainable description of the chosen distractors in this study are as follow:

Table 6. The Classification of Distractor Effectiveness of Test Item on English Final Test

No	Functioning	Total	Test Item	Percentage	Category
1.	4	8	3, 9, 18, 20, 27, 35, 36, 40.	20%	Very Good
2.	3	17	1, 2, 7, 8, 10, 11, 12, 14, 19, 24, 25, 26, 29, 30, 33, 38, 39.	43%	Good
3.	2	11	4, 5, 6, 13, 16, 21, 22, 23, 31, 32, 37.	27%	Fair
4.	1	4	15, 17, 28, 34.	10%	Less Good
5.	0	0	-	0	Not Good
Total		40		100%	

The table above presents the data about the detail description about the quality of test item based on the chosen distractor effectiveness of test item on English final test for the tenth grade of SMK PGRI Wonoasri. There are eight items or 20% of the total test items that are classified as *very good* items because all the distractors in each item are functioning. Those items which belong to this category are test item number 3, 9, 18, 20, 27, 35, 36, and 40. For the category of *good* items, there are seventeen or 43% of the total test items which become the highest amount within the categories. Test items that identified as *good* items are test items number 1, 2, 7, 8, 10, 11, 12, 14, 19, 24, 25, 26, 29, 30, 33, 38, and 39. Afterwards, the test items number 4, 5,

6, 13, 16, 21, 22, 23, 31, 32, and 37 are categorized as *fair* items because there is only one functioning distractors in each of the test item. The percentage of this category is 17% with the total amount eleven items. The next category is *less good* items. This category has the lowest amount among the other categories with the total amount four items and 10% as the percentage. Items that appears within this category are test item number 15, 17, 28, and 34 and recommended for revision on the non-functioning distractors. In the last category, there is none of the test item that appears as *not good* items. In this case, it can be concluded that most of the test items on English final test based on the distractor effectiveness' recommendation are good items. This result is very contradictive with the previous research conducted by Hartati and Yogi (2019) in which none of their distractors were effective.

There are other categorization of test item based on functioning and non-functioning distractors on multiple choice questions. Those categories are *good*, *fair*, and *less good* items. These categorization is in accordance with the previous research conducted by Herlambang in 2015. He framed on test item into *good*, *fair*, *less good*, and *not good* item based on the functioning and non-functioning distractor. In the same year, Haryudin (2015) and Manfenrius et al (2015) reported their study related to the effectiveness of the distractors in their item test. The result shows that more than 40% were ineffective.

The overall results of item analysis on multiple choice questions of English final test for the tenth grade of SMK PGRI Wonoasri based on difficulty level, index discrimination, and distractor effectiveness have been illustrated above. Nevertheless, the summary of the whole results analysis are as follow:

Table 7. Overall Results of Item Analysis on Multiple Choice Questions of English Final Test for the Tenth Grade of SMK PGRI Wonoasri

No.	Criteria	Test Item	Total	Percentage	Note
1.	3 criteria	1, 3, 9, 18, 19, 29, 30, 32, 33, 36, 37, 38.	12	30%	Accepted
2.	2 criteria	2*, 4, 5*, 6, 7*, 8, 10, 11, 12, 13*, 14, 16, 20*, 22, 23, 24, 25*, 26, 27, 31, 35, 39, 40.	23	57%	Revision
3.	≤ 1 criteria	15, 17, 21, 28, 34.	5	13%	Not Accepted

*) negative index on the level of discrimination

Table 7 above illustrates the data regarding to the overall results of item analysis on multiple choice questions of English final test for the tenth grade of SMK PGRI Wonoasri based on difficulty level, index discrimination, and distractor effectiveness. From the data above, it can be seen that the results of the analysis is obtained by the some criteria. Those criteria are: (a) test item is considered for being accepted if the test item has met of three criteria; (b) test item is considered for revision if the test item has fulfilled of two criteria; (c) test item is considered as not accepted or rejected if the test item only fulfilled by less or one criterion;

The next finding based on the data displayed, for the first, it shows that there are 12 items (30%) which are accepted. It means that these items are good items which can be saved in the bank of questions and can be used for the future test. These good items are test item number 1, 3, 9, 18, 19, 29, 30, 32, 33, 36, 37, and 38.The second phenomenon, it can be seen that there are 23 items (57%) which become the highest amount, is needed for revision. Nevertheless, among those 23 items, there are six items with negative results on discrimination index. Thus, these six items which are item test number 2, 5, 7, 13, 20, and 25, should be replaced with a new one. In this case, there are seventeen items (42%) that need for revision and six items (15%) should discharged and replaced.

For the last criterion, not good items, which become the smallest amount among those three criteria, reveals five (13%) items that include in it. Those five items are test item number 15, 17, 21, 28, and 34. In this case, the items that belongs to not accepted items are categorized as bad items and suggested to be discharged or replaced with a new one.

In conclusion, test items on multiple choice questions of English final test for the tenth grade of SMK PGRI Wonoasri are mostly categorized as *fair* items because most of the items which are 42%, are need to be revised. However, there are still a lot of test items which are 38%, are bad items and need to be discharged. The reason why does the test item is recognized as a bad item can be seen through the table below:

Table 8. The Cause of Bad Items on Multiple Choice Questions of English Final Test

No	The Cause of Bad Items	Test Item/ Question Number	Total (item/s)
1.	Difficulty Level	12, 15, 16, 17, 21, 28, 34.	7
2.	Discrimination Index	2*, 4, 5*, 6, 7*, 8, 10, 11, 13*, 14, 15, 20*, 21, 22, 23, 24, 25*, 26, 27, 28, 31, 36, 39, 40.	24
3.	Distractor Effectiveness	15, 17, 28, 34.	4

*) negative discrimination index

Table 8 shows the data about the cause of bad items on multiple choice questions of English final test. It can be noticeable that from the table above, the highest number that cause an item being bad is the index of discrimination level. There are twenty four items, including the negatives results which having a low index. In this case, those items that cause by the discrimination index are not able to distinguish the students. In this terms refers to distinguish between smart students and the less one. While distractor effectiveness as the lowest amount that cause an item for being categorized as bad items with four items included with in it, is less affecting because of the significant gap of amount. All in all, these bad items either can be revised or discharged as the solutions.

4. CONCLUSION

Based on the result of test item analysis that involves difficulty level, discriminating index, and distractor effectiveness on multiple choice questions of English final test for the tenth grade of SMK PGRI Wonoasri, it can be discovered that the result of this study shows: (1) the level of difficulty which categorized as difficult is 35% (40 items), medium is 47% (19 items), and easy is 18% (7 items); (2) discrimination index that makes the test items are recommended for being rejected is 15% (6 items), accepted 38% (15 items), and not accepted is 47% (19 items); (3) distractor effectiveness which has functioning distractors and categorize an item test as very good is 20% (8 items), good is 43% (17 items), fair 27% (11 items), less good 10% (4 items). Taking everything into account, the quality and acceptance of test items on multiple choice questions of English final is medium or fair. For that reason, to produce a good item of multiple choice questions which based on the classic theory measurement of test item analysis, a test item should meet three criteria. Those three criteria are difficulty level, discriminating index, and distractor effectiveness. In closing, there are still a lot of items on multiple choice questions of English final test need to be revised.

Refer to the result of this study, the researchers suggest that: (1) the teacher as a test constructor should knowledgeable about the characteristics of good language test item, especially procedures of determining difficulty levels, and discrimination index and distractor effectiveness; (2) the teacher should finds another methods to teach English so that the students are fully understand on the topic which being stated in the test items; (3) the test items which are suggested for revision, should be saved and pre-tested for further analysis. Test items that contain too many problems should be discarded or replaced with the new one; and (4) the researchers hope that the result on this item analysis able to be used as an example on analyzing the other test items and encourages other researchers to study on the same or similar subject.

REFERENCES

Ali, S.H.Carr.P.A, & Ruit.K.G. (2016). *Validity and Reliability of Scored Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter*. Journal of The scholarship of Teaching and Learning. 16(1).

Arikunto, S. (2016). *Dasar-Dasar Evaluasi Pendidikan* (2nd.ed). Bumi Aksara.

Brown, H.D. (2001). *Teaching by Principle: An Interactive Approach to Language Pedagogy* (2nded).

Brown, J.D. (1996). *Testing in Language Programs*. Prentice Hall Regents, Prentice-Hall, Inc.

Haladyna, T.M. 2004. *Developing and Validating Multiple-Choice Test Items* (3rded). Lawrence Erlbaum Associate Publishers.

Hartati, N.,& Yogi,H.P.S. (2019). *Item Analysis For a Better Quality Test*. English Language in Focus (ELIF).

Haryudin, A. (2015). *Validity & Reliability of English Summative Test at Junior High School in West Bandung*. Jurnal Ilmiah UPT P2M STIKIP Siliwangi.2(1).

Herlambang, Bima Kartika. (2015). *Analisis Butir Soal Ulangan Tengah Semester Mata Pelajaran Pendidikan Jasmani Olahraga Kesehatan Kelas VII Semester Genap SMP N 2 Wonosari Tahun Ajaran 2014/2015*. Universitas Negeri Yogyakarta.

Jannah.R, Hidayat.D.N, Husna. N. Khasbani, I. (2021). *An Item Analysis on Multiple Choice Question: a Case of A Junior High School English Try Out Test in Indonesia*. Jurnal Bahasa, Sastra, dan Pengajarannya. Universitas Muhammadiyah Purwokerto.

Kashap, Surekha. (2015). *Item Analysis of Multiple Choice Questions*. International Journal of Current Research, 7(12).

Longman. Brown, H.D. (2004). *Language Assessement: Principle and Classroom Practices*. Longman.

Leavy, Patricia. (2017). *Research Design; Quantitative, Qualitative, Mixed Methods, Arts-Based, and Community-Based Participatory Research Approaches*. The Guilford Press.

Madsen, H.S. (1983). *Techniques in Testing*. Oxford University Press.

Maharani, A.V., Putro, N.H.P.S .(2020). *Item Analysis of English Final Test*. Indonesian Journal of EFL and Linguistics.5(2).

Manalu, D. (2019). *An Analysis of Students Reading Final Examination by Using Item Analysis Program on Eleventh Grade of SMA Negeri 8 Medan*. JETAL: Journal of English Teaching & Applied Linguistics. 1(1).

Manfenrius,A.,Sutopo,G & Wijaya.B. (2015). *Item Analysis on English Summative Test at the Eight Grade Junior High School in Pontianak.* Jurnal Pendidikan dan Pembelajaran Khatulistiwa, 4(12).