

An Analysis Of Discrimination Power, Difficulty Index, And Distracter Of English Test Items Constructed By Efl Teacher Of Eleventh Grade At Negeri 2 Senior High School Maumere

Kornelia Yosefina Lani¹, Maria Yulita Dhewa², Oktaviana Perada Samon³, Maria Donna Chelsea⁴, Anjelina Nidia De'e⁵, Ana Maria Avelina⁶

¹Nusa nipa university, ²Nusa nipa university, ³Nusa nipa university, ⁴Nusa nipa university, ⁵Nusa nipa university, ⁶Nusa nipa university

Corresponding Author: Kornelia Yosefina Lani, Lani, e-mail: yesinesa24@gmail.com

Published: July, 2026

ABSTRACT

The purpose of this study is to evaluate the discrimination power, difficulty index, and effectiveness of the distractors for the English test items prepared by the EFL teacher of the eleventh grade students at SMA Negeri 2 Maumere. The present study used a descriptive qualitative approach where the objects of the research were 13 English test items and answer sheets of 20 eleventh-grade students. Data collection was done through documentation while manual analysis using item analysis formula was done to find the difficulty index, discrimination power, and distractor index of the items. It is found that concerning the discrimination power, there was one (7.7%) dropped item, seven (53.8%) items were categorized as corrected, and five (38.5%) items were categorized as good. With respect to the difficulty index, three (23.1%) items were easy, five (38.5%) were moderate, and five (38.5%) were difficult. Concerning the distractor analysis, about 12% distractors were very poor, 31% were poor, 25% were fair, 15% were good, and 19% were very good, which implies that many distractors were not effective in drawing the attention of low performing students to the wrong answers. With regards to the above findings, it can therefore be concluded that the English test items formulated by the EFL teacher do not meet the requirements of good test items with regard to discrimination and distractors and thus need further improvement.

Keywords: Test Item Analysis, Distracter, Discrimination Power, Difficulty Index

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

INTRODUCTION

Teaching is correlated with learning. The experimental and theoretical study of learning is an extensively and intensively cultivated branch of psychology. According to (Association, 2020) Teaching is the process of facilitating learning and helping learners acquire knowledge, skills, values, and habits. It means that teaching is an activity that helps students gain knowledge, develop skills, and learn positive values and habits. Through the teaching process, teachers guide and support learners so that they can achieve their learning goals effectively. Moreover, teaching is not just the transfer of information from the teacher to the learner, but it is the interaction between both the individuals through which the learner develops critical thinking and applies his/her acquired knowledge and skills in real life situations. In this regard, a good teacher should know the requirements, strengths and weaknesses, and preferred modes of learning of the learners to develop relevant teaching approaches for them. Teaching is also an ongoing activity in which assessment and evaluation of learners' performance takes place. Thus, the efficiency of the teaching process can be assessed not only on the basis of the way material is conveyed to the learners but also on the basis of the learners' performance.

Teaching English is the process of teaching students to understand and use English for communication and learning purposes. It includes developing language skills such as listening, speaking, reading, and writing, as well as learning grammar, vocabulary, and pronunciation. According to Maqsood et al. (2024) Teaching English in a cultural context means that the approach within language education such as English that interrelated with the cultural aspects and facts of the English language into the teaching and learning process for teachers and learners as well (P.1). It means that teaching English is not only focused on learning language skills, but also on understanding the culture associated with the English language. By integrating cultural aspect into the teaching and learning process, students can develop better communication skills and gain a deeper understanding of how English is used in real-life situations. Also, when teaching English as a foreign language, one needs to keep in mind the various abilities of the learners, their motivation, and the level of their language exposure outside class hours. One needs to create a friendly environment where students can actively engage themselves in learning English and not just memorize rules and lists of words. Different methods of instruction can be employed here to ensure that the entire process becomes more meaningful and realistic. The main purpose of English instruction should not be confined to ensuring that learners pass their exams, but they develop effective communication skills for real-life situations.

Evaluation is an important process because it helps determine the quality and value of something. The concept of evaluation is still debated, and many people find it difficult to distinguish evaluation from research. Therefore, a clear evaluation process is needed to assess whether English test items are appropriate and meet educational objectives. According to Wanzer, (2020) Evaluation is a process that involves determining the value or benefits of a particular program, activity, or object based on the goals and needs of the users of the evaluation results. It means that the evaluation process is not just about measuring outcomes but assessing whether such outcomes are significant and useful for the stakeholders of the process. When it comes to education, evaluation becomes an essential component of the process that allows teachers and educational institutions to assess whether the teaching or learning process has met its goals. Another important function of evaluation is the collection of data that may prove useful in improving the quality of instruction, learning materials, as well as assessment tools, among which there are test items. If there were no evaluation, it would be impossible to find out the advantages and disadvantages of a test, a curriculum, or a teaching technique because all judgments would be made based on assumptions. Moreover, the evaluation process in language testing contributes to ensuring that the test items are valid and reliable and may provide accurate measurements of language abilities of the students.

The measurement consists of language tests which were previously employed to determine whether one has passed the examination or not. Due to this fact, a language test should be viewed as being a good one when done by the test maker. In order to achieve this goal, the test maker has to know how to construct a good test. A good test is supposed to have good test items. Good test items should include some criteria such qualitative and quantitative analysis. A test should be regarded as good when constructed properly by the test maker, and achieving this requires assembling test items that meet established quality criteria, since item analysis is generally understood as the set of qualitative and quantitative procedures used to evaluate the characteristics of test items both before and after test development and construction (Rashwan et al., 2024). It means that a test maker cannot just depend on intuitive skills for developing test items but should use certain procedures so that every item fulfills its purpose. Through qualitative analysis, the test maker is able to decide whether the test items are clear enough, unambiguous and meet the required criteria and guidelines of item development. Quantitative analysis, in its turn, helps to obtain empirical information about how the items work in real conditions when used by students. It includes such aspects as difficulty of the item, ability of the item to discriminate between high-achievers and low-achievers, effectiveness of the distractors, etc. Qualitative and quantitative analyses are two interconnected sides of one coin; thus, none of them can provide full information about items. Qualitative judgment alone cannot guarantee

statistical preciseness, while statistical data alone cannot reveal problems in item development and content validity of the test. Therefore, through qualitative and quantitative analysis, a test maker can distinguish weak items which have to be improved or even deleted and select items that have been proved to be good.

Test item analysis is done to determine dropped, revised and good test items. In order to get information on good test items, these need to get information on good test items, there is need to get information regarding three forms of analysis concerning the item analysis which include level of difficulty, discrimination power, and distracter, A good test items is an item which is neither too difficult nor too easy. An item which is too easy does not motivate the student give up since they are not motivated to do it again because is out of their reaches to finish it. On the contrary, a question that is too difficult may prove demotivating to learners since the perception that however much they work at it they will never succeed in solving the task may reduce their motivation to attempt the test. Thus, a good test question should be developed with a moderate degree of difficulty, one that would be challenging enough to separate the students that have comprehended the content from those who have not, but would still be solvable by those students who have worked diligently. Apart from having an appropriate degree of difficulty, an efficient test item should exhibit a high level of discrimination power that would allow distinguishing between high-performing and low-performing students based on their true abilities and not on guessing. Moreover, the distracters used in multiple choice test items should be created in such a way that they would appear convincing to those students who have failed to grasp the subject, while those who have successfully mastered the content would be able to recognize that they were incorrect. If all these three characteristics are analyzed adequately, then the developed test questions could be regarded as effective assessment tools for evaluating students' linguistic competence.

Based on the background above, this study aims to analyze the discrimination power, difficulty index, and distractors of English test items constructed by the EFL teacher for eleventh grade students at SMA Negeri 2 Maumere. Specifically, this study is intended to: (1) describe the discrimination power of the English test items, (2) describe the difficulty index of the English test items, and (3) describe the quality of the distracters used in the English test items. Through this analysis, the study seeks to determine whether the test items constructed by the teacher meet the standard criteria of a good test item, and to provide useful feedback for improving the quality of English test construction at the school in the future.

THEORICAL FRAMEWORK

The theoretical foundation of the present study rests primarily upon T.L. Kelley (1939) namely Kelley's discrimination index. formulated the early statistical approach for calculating item discrimination, which isolates the upper and lower 27% of a testing group to compute an item's effectiveness. Discrimination index (D) is a classical test theory statistics indicator that is used to measure an item's capability to distinguish between examinees of high and low performances, and which is calculated based on the upper and lower 27% rule by comparing proportions of high-scoring examinees' correct answers with proportions of low-scoring examinees' correct answers. (Rezigalla et al., 2024) This method determines the difference in difficulty between top 27% and bottom 27% examinees, and is additionally supplemented with the correlation coefficient between an item's score and test score, while investigations conducted on real examination data demonstrated that there is a negative correlation between the average difficulty of items and average discrimination indicating that easy items have less discrimination properties than difficult ones. Item discrimination properties were recently shown to be related to item construction it was found that item discriminating index was significantly higher for items with a single distractor functioning correctly rather than for items with three distractors functioning correctly, indicating that discrimination properties are not the exclusive property of the correct answer itself but are also influenced by the distractors.

The difficulty index is defined within classical test theory as the proportion of correct responses obtained per item, with low difficulty indices indicating hard questions and high difficulty indices indicating easy questions (El-Hamamsy et al., 2025). Numerous thresholds have been employed in the literature to identify items that are too easy or too hard, although these thresholds are often arbitrary; items with difficulty values between 4 and 6 are generally considered to have maximum discrimination potential, and commonly used thresholds tend to vary around these values. This is consistent with related work showing that items with difficulty indices above a 0.85 threshold are typically considered too easy, while items with difficulty indices below a 0.25 threshold are considered too difficult, reinforcing the idea that difficulty and discrimination are interdependent properties rather than isolated metrics of item quality.

In an objective test that utilizes multiple choice questions, each question presents several possible answers. These possible answers are commonly called options or alternatives, ranging from four to five in number per question (Sudijono, 2013). Of the various options available for each question, there will be only one right answer and the rest will be wrong answers, commonly called distractors, which means that each item of a five-option test form will consist of one right answer and four distractors (Sudijono, 2011:409). To evaluate the level of distractor computation, the level of each distractor is standardized into five classes, namely very good, good, fair, poor, and very poor, according to certain numerical ranges; a distractor level of 0.76 – 1.25 falls into the very good category; 0.51 – 0.75 and 1.26 – 1.50 fall into the good category; 0.26 – 0.50 and 1.51 – 1.75 fall into the fair category; 0 – 0.25 and 1.76 – 2.00 fall into the poor category, while any level above 2.00 falls into the very poor category (Nitko & Brookhart, 2006).

METHOD

In this research methodology, the researcher elaborated the aspects of research preparation, scope or object of research, instrument used for gathering data, location of the research, research variable, and the technique of data analysis. This research adopted descriptive qualitative research as its method due to the purpose of conducting this research that is describing and explaining the characteristics of the test items made by the EFL teacher concerning discrimination power, difficulty index, and distractor power without testing any hypothesis or generalizing the results. The data collected after calculating the index will be interpreted and described qualitatively based on the criteria. The target of this study was the English test items developed by the EFL teacher for the eleventh graders, as well as the answer sheet completed by the students in taking the test. The test items were used as the main target that would be analyzed based on its level of difficulty, discrimination ability, and distractors, whereas the answer sheets were used as the source of data in calculating all three indices. The study was carried out in SMA Negeri 2 Maumere, which particularly involved the eleventh graders of the school. Data collection in this study was done by documentation, specifically the English test items developed by the EFL teacher and the answer sheets completed by the students of grade eleven. Such documents became the major instruments of this research since they were the basis of the raw data used in calculating the difficulty index, discrimination power, and effectiveness of the distractors of the test items. The variables in this study included discrimination power, difficulty index, and distractor of the English test items developed by the EFL teacher for the eleventh grade students of SMA Negeri 2 Maumere. The data used in this research study have been manually analyzed with the use of appropriate formulas for Item Analysis. First, the difficulty index of each test item was determined in order to find out if the test item is easy, moderate, or difficult. Second, the discrimination power of each test item was found out in order to measure the extent to which that particular test item can distinguish between high achievers and low achievers. Third, the distractor analysis of each test item was performed in order to check the validity of the distractor options in the test items.

There are some steps of analyzing the data that used in this research study. It is described as follow:

1. Counting the level of difficulty using the formula:

Difficulty level or difficulty index

$$DI = \frac{C}{N}$$

DI : Difficulty Index

C : The number of students who have correct answer

N : Number of analyzed students

(Puspendik 2013: 9)

The difficulty index calculated to know the index of difficulty of both groups in the level of easy, moderate or difficult. It standardized as the following:

0 < DL ≤ 0,3 : Difficult

0,30 < DL ≤ 0,70 : Moderate

> 0,70 : Easy

2. Counting the discrimination power using the formula:

$$DP = \frac{2 (HC-LC)}{N}$$

DP : Discrimination Power

HC : The number of students of high group who have correct answer

LC : The number of students of low group who have correct answer

N : Number of analyzed students

(Puspendik 2013:8)

To know the test items is good, corrected or dropped; discrimination power has the certain standard to decide it. We can see in the following formula:

DP ≤ 0 : Dropped

0 < DP ≤ 0,25 : Corrected

DP > 0,25 : Good

3. Counting the distracter using the formula:

$$Dtl = \frac{nDt}{(N-Nc)/(Alt-1)}$$

Dtl : Distracter Index

nDt : Total students who choose the distracter

N : Total test takers

nC : Total correct answer

Alt : Total number option

To know the result of computation of distracter above, it presented the level of distracter as the following table:

Table 1. Distracter range and level	
Distracter Range	Distracter Level
0.76 to 1.25	Very good
0.51 to 0.75 and 1.26 to 1.50	Good
0.26 to 0.50 and 1.51 to 1.75	Fair

0 to 0.25 and 1.76 to 2.00	Poor
More than 2.00	Very poor

FINDING AND DISCUSSION

Findings

In this part, the writer will analyze each test items constructed by EFL teachers at Negeri 2 Maumere Senior High School in the academic year of 2026/2027. There are 13 test items for nine grade students. This analysis aims to examine the quality of the test items, including whether they can distinguish between higher group and lower group performing students, their level of difficulty, and whether the distractor (wrong answer choices) function effectively for both groups. The three aspects of test item analysis are described below:

1.1 Discrimination Power

The discrimination power of the test items constructed by EFL teachers at Negeri 2 Maumere Senior High School in the academic year of 2026/2027 is presented in table 2.

Table 2. Discrimination Power

No	Discrimination Power	Decision
1.	0,10	Corrected
2.	0,50	Good
3.	0,20	Corrected
4.	0,20	Corrected
5.	0,20	Corrected
6.	0,20	Corrected
7.	0,20	Corrected
8.	-0,10	Dropped
9.	0,50	Good
10.	0,30	Good
11.	0,40	Good
12.	0,20	Corrected
13.	0,50	Good

Based on Table 2 of the discrimination power analysis of the 13 test items, there was 1 test item (7.69%) classified as Dropped and therefore could not be used, namely item number 8. In addition, there were 7 test items (53.85%) classified as Corrected or needing revision, namely items 1, 3, 4, 5, 6, 7, and 12. Furthermore, there were 5 test items (38.46%) classified as Good and suitable for use because they had good discrimination power, namely items 2, 9, 10, 11, and 13.

2.1 Difficulty Index

The difficulty index of the test items constructed by EFL teachers at Negeri 2 Maumere Senior High School in the academic year of 2026/2027 is represented in the following table 3.

Table 3. Difficulty Index

No	Difficulty Index	Decision
1.	0,85	Easy
2.	0,35	Moderate
3.	0,80	Easy
4.	0,30	Difficult
5.	0,20	Difficult
6.	0,40	Moderate
7.	0,10	Difficult
8.	0,85	Easy
9.	0,45	Moderate
10.	0,25	Difficult
11.	0,50	Moderate
12.	0,30	Difficult
13.	0,35	Moderate

Based on Table 3 of the difficulty index analysis of the 13 test items, there were 3 test items (23.08%) classified as Easy, namely items 1, 3, and 8. In addition, there were 5 test items (38.46%) classified as Difficult, namely items 4, 5, 7, 10, and 12. Furthermore, there were 5 test items (38.46%) classified as Moderate, namely items 2, 6, 9, 11, and 13. Therefore, most of the test items were in the Difficult and Moderate categories, each representing 38.46% of the total items, while 23.08% of the items were in the Easy category.

3.1 Distracter

The result of distracter analysis of 13 test items construction by EFL teachers at Negeri 2 Maumere Senior High Schools in the academic year 2021/2027. The data are show below:

Item number 1

$$Dtl = \frac{nDt}{(N-nc)/(Alt-1)}$$

- Option A

$$\begin{aligned}
 Dtl &= \frac{3}{(20-17)/(5-1)} \\
 &= \frac{3}{3/4} \\
 &= 3
 \end{aligned}$$

$$\begin{array}{r} \hline 0,75 \\ = 4 \end{array}$$

- Opsion B

$$\begin{array}{r} \text{Dtl} = \frac{0}{(20-17)/(5-1)} \\ = \frac{0}{3/4} \\ = \frac{0}{0,75} \\ = 0 \end{array}$$

- Opsion C

$$\begin{array}{r} \text{Dtl} = \frac{0}{(20-17)/(5-1)} \\ = \frac{0}{3/4} \\ = \frac{0}{0,75} \\ = 0 \end{array}$$

- Opsion D

$$\begin{array}{r} \text{Dtl} = \frac{0}{(20-17)/(5-1)} \\ = \frac{0}{3/4} \\ = \frac{0}{0,75} \\ = 0 \end{array}$$

The test items number 2 until 15 are analysis as example of analysis above. It show on the follow table:

Table 4. Distracter

No	A	B	C	D	E
1.	4	0	0	0	*
2.	*	1,23	0,92	1,54	0,31
3.	*	1	2	1	0
4.	*	1,71	0	0,57	1,71
5.	*	0,75	0,50	0,25	2,50
6.	3	0	*	0,67	0,33
7.	0	2,89	1,11	*	0
8.	0	1,33	*	0	2,67
9.	1,45	0,73	1,82	*	0
10.	*	1,60	1,87	0,27	0,27

11.	2	*	0,80	0,40	0,80
12.	*	0,29	2,57	0,29	0,86
13.	0,92	0,62	1,54	*	0,92

The result of distracter analysis above is interpreted by using standardize level of distracter that is presented in the following table 5.

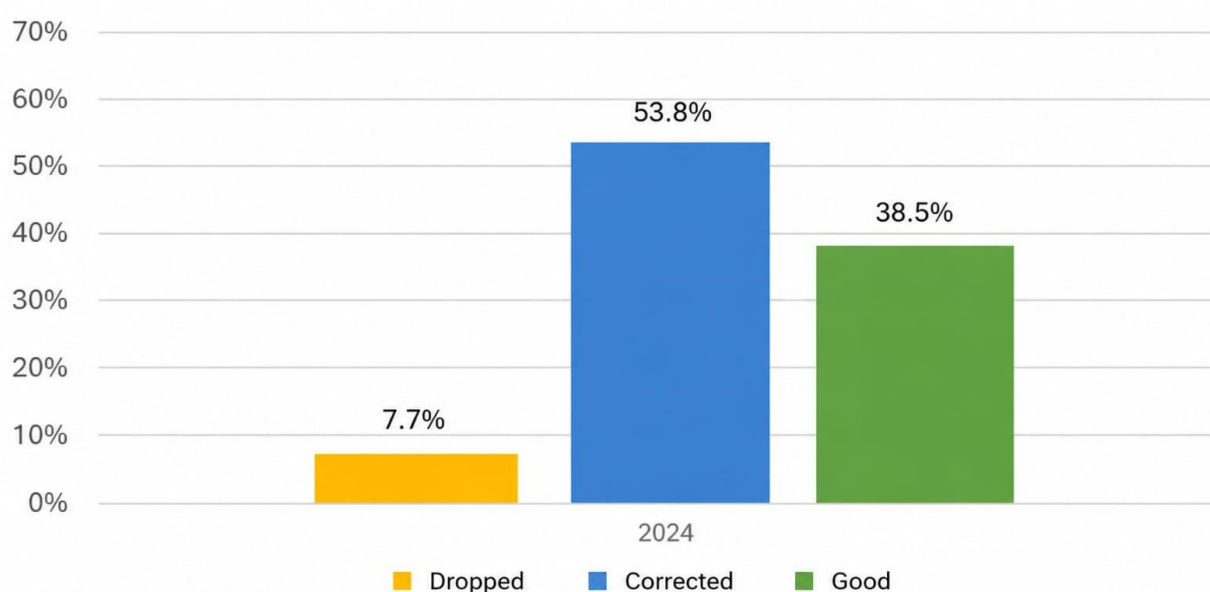
Table 5. Distracter Level

Level	A	B	C	D	E	Σ	%
Very poor	2	1	1	0	2	6	11,54
Poor	3	2	5	3	3	16	30,77
Fair	0	3	2	4	4	13	25,00
Good	1	4	0	2	0	7	13,46
Very good	1	2	3	1	3	10	19,23
	7	12	11	10	12	52	100

Based on Table 3.3.1 of the distracter level analysis, there were 6 distracters (11.54%) classified as Very Poor, 16 distracters (30.77%) classified as Poor, 13 distracters (25.00%) classified as Fair, 7 distracters (13.46%) classified as Good, and 10 distracters (19.23%) classified as Very Good. Therefore, most of the distracters were in the Poor category with 30.77%, followed by the Fair category with 25.00%. Meanwhile, distracters in the Good and Very Good categories made up only 32.69% of all distracters.

Discussion

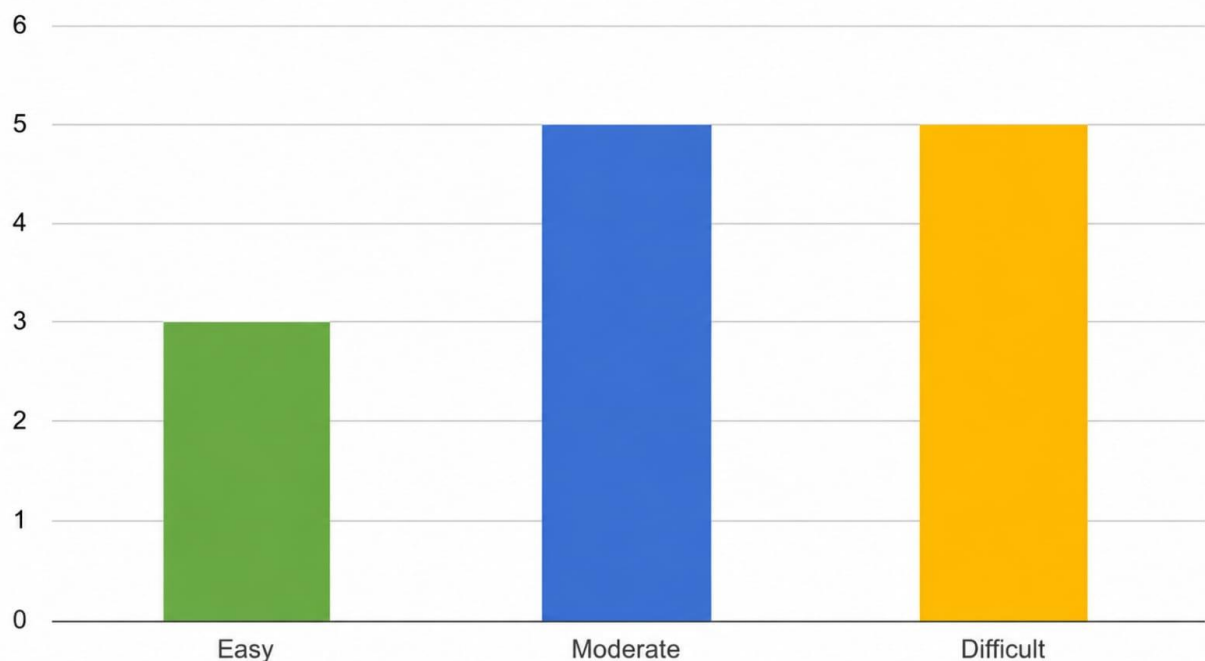
Figure 1. Discrimination Power



From the results based on Chart 4.1 on discrimination power analysis on the test items of English constructed by the EFL teacher of eleventh grade at SMA Negeri 2 Maumere, out of 13 items,

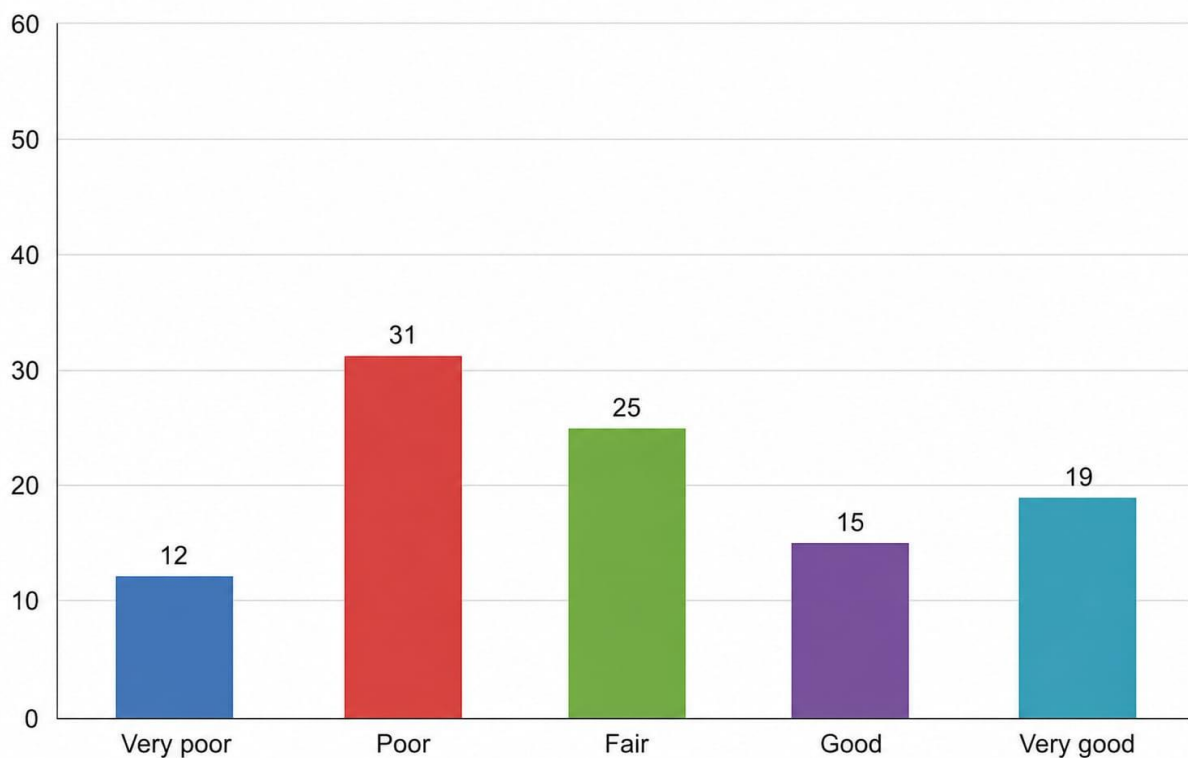
1 item (7.7%) was classified as dropped, 7 items (53.8%) were classified as corrected, and 5 items (38.5%) were classified as good. These findings imply that more than half of the test items did not yet have discrimination power and thus need to be revised, while only few test items were effective in discriminating high-achievers from low-achievers. These results coincide with another study performed on Grade X English test items developed in collaboration with an English teacher in a private high school in Manado City, North Sulawesi (Rezkilaturahmi & Rosnawati, 2024) where out of 50 items, only 11 items were found qualified based on their difficulties and discrimination power, while other items contained 1 to 3 distractors that did not function well. The two studies similarly show that test items constructed by EFL teachers, when tested using classical test theory, tend to produce low percentage of test items with high discrimination power.

Figure 2. Difficulty Index



The analysis of the difficulty index of the English test items developed by the EFL teacher for the eleventh grade of SMA Negeri 2 Maumere using Chart 4.2 revealed that among 13 items tested, 3 items were found to be easy, 5 items were classified as moderate, and 5 items were classified as difficult. It can be concluded from this analysis that most of the test items belong to the moderate to difficult category, while few items were found to be easy for the students. A more comparable distribution can also be seen in another study in , where the analysis of English test items was conducted in the academic year 2022-2023 and revealed that among 25 multiple choice items, the item difficulty was considered to be easy, where there were 2 difficult items, 7 medium items, and 16 easy items, but the item discrimination was found to be poor. Contrary to the results in this study, where the majority of items are easy, the results in the current study reveal the opposite trend, where the number of moderate and difficult items is greater than that of the easy items. This implies that unlike the findings in the Gowa study, the test items developed by the EFL teacher in SMA Negeri 2 Maumere were more difficult for eleventh grade students.

Figure 3. Distracter



The distractor analysis of the test items for English prepared by the EFL teacher of eleventh grade at SMA Negeri 2 Maumere based on the chart above found out that the distractors were categorized into five levels of quality, namely, very poor, poor, fair, good, and very good where poor is the largest group, followed by fair, very good, good, and then very poor being the smallest. The result implies that many of the distractors in the test items were not effective in diverting low-ability students from choosing the right answers. This result is supported by research done by (Rejeki et al., 2023) where they did a small-scale study of the quality of the teacher's summative test and discovered that 43 distractors had to be revised in order to increase their discriminating ability. In this case, this study also supports such an idea because, since the high presence of items belonging to "poor" and "fair" groups, there should also be a need to revise the distractors made by the teacher at SMA Negeri 2 Maumere. A wider comparative discussion may also be made between this paper and another research involving an analysis by (BINDOWO & IBRAHIM, 2025) where 59 of the distractors examined were said to have been defective and unable to satisfy the set minimum standards, besides poor discrimination power noted in most of the test questions. Despite the fact that this latter Nigerian study was carried out under a different country and assessment framework, the general trend observed in the study is similar to that observed in the current study. There is a high possibility that many teacher-made or exam board-made distractors will not work properly, whatever the particular education system. Such comparison indicates that the issues related to the quality of test items found through the analysis of distractors for English test items in SMA Negeri 2 Maumere are not unique. As mentioned above in relation to the distractor analysis, by conducting a distractors analysis one can improve the items under examination and make them clear and relevant to the learning objectives, and using often chosen distractors teachers can adjust their teaching methods in accordance with the difficulties of the students. Thus, distractors classified as poor or very poor in the current research should not be discarded, but on the contrary, used for further improvement of both test items and teaching methods applied by the EFL teacher to overcome the students' misconceptions.

CONCLUSIONS

Based on the results and discussions discussed above, it can be stated that the test items constructed by the EFL teacher of eleventh grade at SMA Negeri 2 Maumere have not yet been able to meet completely the criteria of good test items in the perspective of three aspects of item analysis including discrimination power, difficulty index, and distractor effectiveness. First of all, in relation to the discrimination power, based on the analysis, it is known that out of 13 items, only a few items were classified as good, many items were classified as corrected, and even a few of them were classified as dropped, thus meaning that most of the items have not yet been able to differentiate clearly between high and low achieving students. The same thing applies to the difficulty index where the items were spread over easy, moderate, and difficult categories and mostly classified as moderate to difficult, which means that the whole test was biased towards a difficult condition. Another aspect of the analysis related to the distractor effectiveness where most of the distractors were found to be classified as poor and fair.

In conclusion, based on these results, it can be stated that although the English test items prepared by the EFL teacher have certain strengths, there are still some aspects of them that must be improved in order to create an accurate and fair measure of students' English language ability through these test items. In fact, these results are similar to the results found in other studies conducted in the same setting as this study; that is, the use of teacher-made English tests. This is because many researchers have shown that teacher-made English test items require systematic item analysis and revision before they become effective measures of language ability. Hence, this study proves the necessity of item analysis as part of test construction, so that teachers would know which test items are good and which are not.

REFERENCES

- Austin, J. L. (1962) *How to Do Things with Words*, Cambridge, Mass.: Harvard University Press
- Bloom, L. and Lahey, M. (1978). *Language Development and Language Disorders*. New York: Wiley.
- Braathe, H. J., & Otterstad, A. M. (2014). Education for all in norway: Unpacking quality and equity. *Procedia-Social and Behavioral Science*, 116, 1193-1200. <https://doi.org/10.1016/j.sbspro.2014.01.368>
- Ghozali, Imam. (2011). *Syllabus and Reading Materials Sociolinguistics*. Yogyakarta.
- Lahey, M. (1988). *What is language? In Language Disorders and Language Development*. London: Collier Macmillan.
- Grice, Paul H. (1975). *Logic and conversation*. In Cole, P., and J.L. Morgan, eds. *Speech Acts*. New York: Academic Press, 41–58)
- Association, A. P. (2020). Publication manual of the American psychological association 2020. *American Psychological Association*.
- BINDOWO, M. G., & IBRAHIM, S. H. (2025). ITEM ANALYSIS OF THE 2019 ENGLISH LANGUAGE MOCK EXAMINATION: IMPLICATION FOR TEST DEVELOPMENT IN TARABA STATE, NIGERIA. *International Journal of Library Science and Educational Research*.
- El-Hamamsy, L., Zapata-Cáceres, M., Martín-Barroso, E., Mondada, F., Zufferey, J. D., Bruno, B., & Román-González, M. (2025). The Competent Computational Thinking Test (cCTt): A Valid, Reliable and Gender-Fair Test for Longitudinal CT Studies in Grades 3–6: L. El-Hamamsy et al. *Technology, Knowledge and Learning*, 30(3), 1607–1661.
- Maqsood, M., Zahid, A., Asghar, T., & Farooqi, S. (2024). *Issues In Teaching English In A Cultural Context*. 4, 1020–1027.
- Nitko, A. J., & Brookhart, S. M. (2006). *Educational assessment of students*. Prentice-Hall, Inc.
- Rashwan, N. I., Aref, S. R., Nayel, O. A., & Rizk, M. H. (2024). Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: implications for developing a validated question Bank. *BMC Medical Education*, 24(1), 168.
- Rejeki, S., Sari, A. B. P., Sutanto, S., Iswahyuni, D., Yogyanti, D. W., & Anggia, H. (2023). Discrimination index, difficulty index, and distractor efficiency in MCQs English for academic purposes midterm test. *Journal of English Language and Pedagogy*, 6(1), 1–11.

- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaiḥi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(1), 445.
- Rezkilaturahmi, R., & Rosnawati, R. (2024). ANALISIS EXPLORATORY FACTOR ANALYSIS (EFA) PADA STRUKTUR TES BAHASA INGGRIS. *JURNAL EDUCATION AND DEVELOPMENT*, 12(3), 70–76.
- Sudijono, A. (2013). *Pengantar evaluasi pendidikan*.
- Wanzer, D. L. (n.d.). *What is evaluation? Perspectives of how evaluation differs (or not) from research*.