LedTure
Journal on Language, Education, Literature and Culture Vol.2, No.1: July 2024
https://bing.ukaw.ac.id/ejournal/index.php/ledture. e-ISSN Online: 3026-0302 DOI: 10.1357/l.v2i1.45

ITEM ANALYSIS OF MULTIPLE-CHOICE QUESTIONS: AN ASSESSMENT ON ENGLISH SUMMATIVE TEST

Gracela Rame[1,], Festif Rudolf Hoinbala[2*]
[1,2]Universitas Kristen Artha Wacana

[*)]Corresponding email: festifrudolf@gmail.com
Received date:24/07/2024; Accepted date:30/07/2024

Abstract: This study aimed to analyze the test items of multiple choice questions on an English summative test. It examines the quality of the teacher-made test in terms of difficulty level, discrimination power, and the effectiveness of distractors. A descriptive method is applied to describe and analyze the data. The instrument for the data collection is a documentary with a document analysis technique in examining the data. The research finding shows that the test has low reliability. Regarding difficulty level, most items (82,5%) are easy, 5 % desirable and 12,5% difficult. Only 7,5 % of the items were categorized as good discrimination power, while 72% were at a poor level and 20% were at a satisfactory level. Furthermore, only 3 % of the questions showed that the distractors worked effectively. It is concluded that the test needed to be revised, and test makers' comprehension of the quality of the test is needed to determine the quality of a reasonable and appropriate language assessment.

Keywords: **test item analysis; difficulty level; discrimination power; distractor efficiency**

INTRODUCTION

A test is a set of questions, exercises, or other instruments designed to gauge a person's or group's level of ability, knowledge, intelligence, abilities, or talent. Arifin (2014) defines a test as a methodology or method used to carry out measurement activities in which there are various questions, statements, or a series of tasks that must be completed or answered by students to measure various aspects of their behavior. In addition, Gampper (2013) adds that the test should be scheduled consistently, generally with a set time, and administered at recognized points in the curriculum. This implies a connection between test design and the actual curriculum used in classrooms. Moreover, according to Ngatman and Fitria (2017), a test is an instrument to gather data about people or things.

There must be a section of a test referring to the qualities of the test. Many experts define some of the qualities of a good test using various criteria. The qualities of the test can be achieved when the items fulfill the criteria of a good test item. Osman (2010) claims that the following factors are used to determine what makes a good test: a) test objectivity, b) test discrimination, c) comprehensiveness, d) validity, e) reliability; f) definition of administering conditions; and g) reliability In addition, Michigan State University Board of Trustees and Omerad (2011) explained that qualities of all good tests are purposeful, valid, reliable, objective, comprehensive, differentiating, expected, instructive and valuable. From those sources, it can be concluded that determining a

test as a good test can be defined by some criteria. It depends on the test evaluators' or test analyzers' criteria.

A variety of test items are used to gauge pupils' performance. The multiple-choice, true-false, completion, matching, short answer, and other objective test elements appear well-known to students. Essay exam items are another common type that allows students to respond to questions in their own words or ideas. The most common type of test used worldwide is multiple choice questions. Although its application has several benefits, multiple-choice questions can be problematic if poorly designed and developed (Marsevani, 2022). Therefore, teachers must have the skills and expertise to design the questions correctly.

The test item is a unique task that the test-taker is required to complete. Sudjana (2014) says that item analysis is the process of examining test questions to produce a set of questions with sufficient quality. Meanwhile, Arikunto (2012) states that analyzing the question items is a systematic procedure that will provide particular information to the prepared test items. Therefore, analyzing the items is very useful for a teacher to know how to make good items. Some procedures should be followed to analyze the test items. These procedures include determining the difficulty level, determining the discrimination power, and analyzing the distractor's effectiveness.

Item analysis is dissecting a single test item to determine and enhance the item's quality (Arikunto, 2005). In addition, Boopathiraj and Chellamani (2013) state that item analysis is carried out to see if the instrument's items belong there. Each item is evaluated for its capacity to distinguish between participants with high and low total scores. According to Thoradik and Hage (1997), examining the test item serves two essential functions. First, it offers diagnostic data for examining the class's learning and its failures to learn, as well as for directing additional instruction and research. Furthermore, Ahmann and Glock (1971) provide a succinct concept in the following passage: Item analysis reviews each test item again to identify its advantages and disadvantages.

According to Widoyoko (2014), there are some reasons why analyzing test items is necessary. They are: (a) determining the test items' strengths and weaknesses, identifying the good ones, or figuring out which ones need to be updated, (b) creating comprehensive information about the specifications of test items to aid teachers in creating tests for any learning evaluation, (c) determining specific test errors, such as answer critical errors, test difficulty levels, and test discrimination abilities. After all, the decision regarding the incorrect test items will soon be made by the teachers acting as the test creator, and (d) the test creator may benefit from a studied test. In this method, the test that has been analyzed can be saved and used as a guide for future test design. Practical exam questions can also be utilized to test a group of students later. There are three main components of item analysis: item facility (also known as item difficulty), item discrimination power (also known as item differentiation) and distractor effectiveness.

A good test item should contain all three levels of difficulty: easy, moderate, and severe. A valid and reliable test should have items from a moderate level. The integrity

of the test may be jeopardized by something that is either too basic or too complex, making it challenging to acquire accurate data on students' academic development. The difficulty index is between 0,00 to 1,0, meaning 0,0 is too tricky, and 1,0 means too easy. The following is the Item of difficulty index (P) according to Arikunto (2005):

Table 1. Criteria of Difficulty Level

Discrimination Power	**Interpretation**
P 0,00	Very difficulty
0,00 < P ≤0,30	Difficulty
0,30 < P ≤ 0,70	Desirable
0,70 < P ≤ 1,00	Easy
P ≤ 1,00	Very easy

The ability of a test to distinguish between proficient pupils and those who are not is known as its discrimination power (Arikunto, 2005). In addition, Harris (1969) defined item discrimination or discrimination power as the ability of a test item to distinguish between test takers with high skill levels and those with low skill levels (Sabri, 2013). A good test question is one that only intelligent pupils can accurately answer. On the other hand, if both high- and low-level pupils correctly answered a question, both levels received credit for the answer. If one is unable to do it correctly, the test item is unreliable.

Table 2. Criteria of Discrimination Power

Discrimination Power	**Interpretation**
DP ≤ 0,00	Very Poor
0,00 < DP ≤ 0,20	Poor
0,21 < DP ≤ 0,40	Satisfactory
0,41 < DP ≤ 0,70	Good
0,71 < DP ≤ 1,00	Very Good

Another crucial indicator of a multiple-choice item's worth in a test is its distractor efficiency, linked to item discrimination. To sum up, the effectiveness of distractor analysis provides data on how effectively a distraction has diverted students who have not done well in their studies from the proper response (Arifin, 2014).

To assess the effectiveness of the tests, the level of difficulty, discrimination power, and distractor efficiency of the ones that have been collected are analyzed. Depending on the following factors, distinguish between the questions that are of good quality, good enough, and not good: (1) If a question satisfies the three requirements of difficulty, distinguishing power, and deceptive efficacy, it is considered to be of high quality, (2) If the question only meets two of the three criteria, it is nevertheless of reasonably good quality, (3) If a question does not satisfy two or all of the requirements, it is considered to be of poor quality.

The fact that the multiple-choice test item is the primary test type used in Indonesia and teachers need to construct a good test encouraged the writers to

research the multiple-choice test item designed by an English teacher. This study aims to determine the difficulty level, the discrimination power and the effectiveness of the distractors in the multiple-choice test made by an English teacher at SMA Negeri 2 Kupang.

METHOD

This study employs a descriptive quantitative technique to describe the quality of the test item in the English subject final test of ten grade students at SMA Negeri 2 Kupang. The instrument serves as a tool for collecting the required data. In this research, the instruments used were the documentation techniques using the English subject final test. The research location was at SMA NEGERI 2 KUPANG, address S.K. LERIK, Kelapa Lima, Kec. Kelapa Lima, Kota Kupang Prov. Nusa Tenggara Timur.

There were 40 questions for the English Summative test, and 36 samples of students' worksheets were used as documentation. Arifin (2012) said there are several steps to analyze test items. First, score all of the students' answer sheets. Then, the scores are recorded, ranging from highest to the lowest. Next, 27% of higher and 27% of lower performers are grouped. At the same time, the medium performers are put aside. Finally, students' answers are analyzed. In this study, an item analysis reveals three things: how difficult each item is, whether or not the question discriminates or tells the difference between high and low students, and which distractors are working as they should. Heaton's (1988) formulas were used to find the level of difficulty and the discrimination power of each item.

The formula used to find the level of difficulty is as follows:

$$FV = \frac{\text{Correct u} + \text{Correct L}}{2n}$$

Where :

FV : Facility value; Level of Convenience
U : The number of correct answers from the upper group
L : The number of correct answers from the lower group
n : The number of all students taking the test

The formula used to find the discrimination power is as follows:

$$DP = \frac{\text{Correct u} - \text{Correct L}}{N}$$

Where:

DP : Discriminating power
U : Sum of students from the upper group who answer correctly
L : Sum of students from the lower group who answer correctly
N : Number of the test-takers in one group

Arifin's (2014) formula was applied to analyze the effectiveness of the distractors.

The formula is as follows:

$$IP = \frac{P}{(N-B)/(n-1)}\ x100$$

Where:

IP : Distractor Efficiency index
P : the number of students who choose deception
N : number of students taking the test
B : The number of students who answered correctly on each question
n : Number of alternative answers (option)

RESULTS AND DISCUSSION

Difficulty Level

The difficulty level is the chance of answering a question correctly at a specific student's ability level. Good questions have a level of difficulty, which is, in a sense, not too easy or too difficult.

Table 3. The results of the difficulty level

Category	Total	Percentage
Very Difficult	3	7,5%
Difficult	2	5%
Desirable	2	5%
Easy	16	40%
Very Easy	17	42,5%

The final exam questions for the English Subject Final Test of Ten Grade Students at SMA Negeri 2 Kupang yielded results on the difficulty of the questions based on the data above. Of these, three or (7,5%) were very difficult questions, two or (5%) difficult questions, 2 (5%) desirable questions, 16 (40%) questions were easy and 17 (42,5%) very easy.

Discrimination Power

A test's discrimination power is its ability to distinguish between students with lower and higher skill levels. A good item must be able to tell students' abilities apart. Description of the results of the analysis discrimination power of the English Subject Final Test of Ten Grade Students at SMA Negeri 2 Kupang: the researchers put it into the classification score that can be seen in Table 4.2 below:

Table 4. The Result of Discrimination Power

Category	Total	Percentage
Very Poor	1	2,5%
Poor	28	70%
Satisfactory	8	20%
Good	3	7,5%
Very Good	0	0%

Based on the data above, the results for the difficulty of the questions are obtained from the English Subject Final Test of Ten Grade Students at SMA Negeri 2 Kupang there are 1(2,5%) very poor, 28(70%) poor, 8(21%) satisfactory, 3(7,5%) good, 0 very poor.

Distractor Efficiency

Malau-Aduli and Zimitat (2012) claim that when distractors fail to attract examinees to choose them, they are dysfunctional and do not contribute to the aim of the assessment. Furthermore, Arikunto (1986) stated that a distractor is considered adequate if it is chosen by at least 5% of test takers.

Table 5. The result of the Effectiveness of Distractor

Category	Total	Present
Effective	4	10%
Less effective	24	60 %
Ineffective	9	22,5%
Dysfunctional	3	7,5%

It can be seen from Table 5 that the distractors in the English test provided for students are only effective for four questions, while sixty percent of the whole questions have less effective distractors. The results also show that three questions have

dysfunctional distractors and nine with ineffective distractors. From this data, teachers should pay attention to providing effective distractors for this type of test.

CONCLUSION

This research highlights three significant findings: level of difficulty, discrimination power, and effectiveness of distractors. Overall, the analysis of the items can be categorized as efficient for the difficulty level and the discrimination power, with a recommendation of revision for some items. However, the design of the distractors in each most important item of the test needs to be reconstructed so they can be used effectively. The results of this study encourage teachers to evaluate the items that are categorized as poor for a better implementation of assessment.

REFERENCES

Ahmann, J. S. & Clock, M. D.1971. *Evaluating pupil growth*. (4th ed.) Boston : Allyn and Bacon, Inc.

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

Arifin, Z. (2014). Evaluasi Pembelajaran. Bandung: PT Remaja Rosdakarya

Arikunto, S. (1986). *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Arikunto, S. (2012). Dasar – Dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.

Arikunto, S. (2013). Prosedur Penelitian: Suatu Pendekatan Praktik. Jakarta: Rineka Cipata

Boopathiraj, C. and 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 Vol.2 (2), 189–193

Gampper, C. (2013). Improving English Test Qualities (pp. 73–83). Thammasat Review, Special Issue

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

Heaton, J. B. (1988). *Writing English Language Tests: A Practical Guide for Teachers of English As a Second or Foreign language.* London:Longman

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer Review Improves the Quality of MCQ Examinations. Assessment & Evaluation in Higher Education, 37(8)

Marsevani, M. (2022). Item Analysis of Multiple-Choice Questions: An Assessment of Young Learners. English Review: Journal of English Education Volume 10, Issue 2, (401-408)

Ngatman & Fitria D. A. (2017). Tes dan Pengukuran untuk Evaluasi dalam Pendidikan Jasmani dan Olahraga. Yogyakarta: Fadilatama

Michigan State University Board of Trustees and Omerad (2011). Handbook of Learner Evaluation & Test Item Construction. Michigan State University

Osman, R. M. (2010). *Educational Evaluation and Testing*. African Virtual University Press.

Sabri, S. (2013). Item Analysis of Student Comprehensive Test for Research in Teaching Beginner String Ensemble Using Model Based Teaching Among Music Students in Public. International Journal of Education and Research, 1(12), 1–14.

Sudjana, N. (2014). Penilaian Hasil Proses Belajar Mengajar. Bandung: PT Remaja Rosdakarya.

Thorndike, R, L. and E. P. Hagen. (1997). *Measurement and Evaluation in Phycology and Education 4th Edition.* Canada: Jon Wiley & Sons.

Widoyoko, E. P. (2014). *Penilaian Hasil Pembelajaran di Sekolah (Cetakan Pertama).* Yogyakarta: Pustaka Pelajar