

An item analysis on multiple-choice questions: a case of a junior high school English try-out test in Indonesia

Rohmatul Jannah¹, *Didin Nuruddin Hidayat¹, Nida Husna¹, Imam Khasbani²

¹Department of English Education, UIN Syarif Hidayatullah, Jakarta, Indonesia

²School of Education, University of Bristol, United Kingdom

ABSTRACT

The present study aims to analyze multiple-choice questions obtained from a trial testing conducted in a state junior high school in Indonesia. The study seeks to reveal the level of difficulty, discriminating power and distractor efficiency of the selected test items by employing item analysis. The result of the study discovers that levels of difficulty on the question items are varied. Some question tended to be easy and moderately difficult while the others are difficult to answer. It also uncovers that, in regard to discriminating power, some questions are well constructed while the others are ambiguously worded that can potentially cause the questions to fail to evaluate the students' ability. The analysis on distractor efficiency presents information how the chosen multiple-choice questions were frequently constructed with less effective distractors that caused more high achieving students to choose wrong answers.

Keywords: *discriminating power, distractor efficiency, item analysis, level of difficulty, multiple-choice questions*

Article History. Received: 8 Oct. 2020, Final revision: 6 Feb 2021, Published: 28 Feb 2021.

Introduction

It is commonly agreed that testing plays a salient role and is of evident importance in the educational system. It has been widely acknowledged that in order to ensure that teaching and learning function properly, testing is needed as it provides both educational authorities and actors with necessary data and documentation (Crocker, 2019). In language education, testing holds a significant position and has been extensively utilized to gauge students' ability and knowledge of the target language. Although the focus of language test appears to change from time to time in response to the shifting perception on what language learners should be able of doing by the end of learning processes (Purpura, 2017), most, if not all, language educators rely exclusively on testing to tell if their teaching has been successful or to ensure if their goals have been met in classrooms. Various course of actions on how to construct test questions have been widely implemented by language educators. Frost (2005) elaborates on various types of testing questions such as: error correction, open-ended questions, true or false, and multiple choice. Among the aforementioned questions commonly used in

language testing, multiple-choice appears to be the most utilized testing items inside classrooms (Bachman & Palmer, 1996; Hemmati & Ghaderi, 2014; Jayanti et al., 2019).

As one of the most prevalent forms of question, multiple-choice is widely used inside classrooms for some reason. As stated by Harris (1969), due to its nature that is "highly structured" (p. 7), multiple-choice question enables test developers to test a wide range of English skills and sub-skills. Despite his critical disposition on the ability of multiple-choice questions to measure English productive skills like speaking and writing, McNamara (2000) agrees that this form of question can serve multiple language testing purposes. Test developers can gauge levels of almost all language skills by using multiple-choice tests. What is more, many experts and researchers (e.g., Javid, 2014; Klufa, 2015; Madsen, 1983; McNamara, 2000) believe, regarding the efficacy of a test, that multiple-choice question does not require much time and efforts to construct, unlike some other types of questions. Multiple-choice questions also provide teachers with ease and flexibility to score the test. Multiple-choice tests

*Corresponding author: Didin Nuruddin Hidayat, Hastuti Townhouse Jalan Keang Risin II RT002/01 No. 116D Legoso Pisangan Ciputat, Tangerang Selatan, Banten, Indonesia 15419. e-mail: didin.nuruddin@uinjkt.ac.id

<https://dx.doi.org/10.30595/lks.v15i1.8768> © 2021 Leksika. All rights reserved.

usually come with a fixed key answer, and scoring can be done with speed and objectively without any test raters' bias (Weimer, 2018). In addition, scoring multiple-choice tests can automatically be done using a machine and, thus, increasing the time efficiency.

The quality of a multiple-choice test is defined by the level of its validity, reliability, and discrimination ability. The validity of multiple-choice test refers to the test's ability to measure what the test is proposed to measure (Dulger & Deniz, 2017). On the other hand, reliability indicates the consistency of the test to measure the test taker ability (Ali et al., 2016). A multiple-choice test is claimed to be reliable when taken by the same person several times, and there is no notable differentiation in the scores obtained. The discrimination ability refers to the ability of multiple-choice items to set high-performing test takers apart from the lower-achieving counterparts. In reassuring that a multiple-choice item has a good level of validity, reliability and discrimination ability, an analysis of difficulty level, discriminating power and distractor efficiency is usually conducted. As postulated by Brown (2004), difficulty level refers to the level of easiness that a multiple-choice item possess. Ensuring that a test item has an appropriate difficulty level is of importance as it will give a significant effect on the ability of the test to measure the true ability of the test takers. Discriminating power, as described by Zajda (2006), is the ability of test item to categorize the test takers in their actual knowledge level. In this case, a test item with an appropriate level of discriminating power can inform if a test taker is a high- or low-performing individual. Madsen (1983) defines distractors as the incorrect alternatives that test developers provide in the questions. Haladyna and Downing (1993) stated that there are two different distractors; a functioning distractor and a non-functioning distractor. Functioning distractor refers to distractors that are chosen by five or more percent test-takers a usually selected by low-performance test takers. In contrast, non-functioning distractors are distractors that are chosen by less than five percent of examinees.

Studies on analysing the quality of multiple-choice tests have been done over several decades all around the world. In Nigeria, for example, Odukoya, Adekeye, Igbinoba, and Afolabi (2017) conducted an item analysis on the multiple-choice tests obtained from four compulsory courses in one of the private universities in the country. The result of the studies indicated that most of the question items were not constructed in such a way

that met the appropriate level of difficulty and good distractive index. The study then suggested the reconstruction of the question items so that they can provide an actual measurement of test takers' skills and knowledge. In a similar vein, Toksöz and Ertunç (2017) carried out an item analysis on multiple-choice items on grammar, vocabulary, reading subjects in one of state universities in Turkey to reveal their difficulty level, discriminating power and distractor efficiency. The outcome of the study showed that the majority of the test items had a moderate level of difficulty while some questions were found to be easy and only few items were regarded to have a high level of difficulty. The study claimed that the proportion of the difficulty level on the test items were appropriate as they would not put a lot of stress on the low-performing students or demotivate the high-achieving ones. In regards to discriminating powers and distractor efficiency, the study found that the test items failed to show an appropriate discriminating power and the distractors used in the questions were less efficient. Thus, to avoid negative washback, the study suggested to moderate the question items. In Indonesian contexts, the study on analysing multiple choice items have been done by some researchers such as: Hartati and Yogi (2019) and Manalu, Sipayung, and Lestari (2019). These studies conducted an item analysis on summative tests obtained from high schools in Indonesia. Just as the previously mentioned studies, these studies focused their attention on discovering the level of difficulty, discriminating power and distractor efficiency from the selected multiple-choice items. The studies yielded different results. Hartati and Yogi, for example, found that the test items from their study were dominated with moderately difficult and easy questions while Manalu, Sipayung and Lestari discovered that the majority of the item questions on their study were moderately difficult and very easy to answer. Concerning the discriminating power, Manalu, Sipayung and Lestari's study uncovered that the most of the test items had a poor discriminating level while Hartati and Yogi's study indicated that the test item had already had an acceptable discriminating level. Both studies however, agreed that there were still some issues on the choice of distraction used on the tests. And for this matter, the studies suggested revision on the wording of the distractors so that they can function better.

While the aforementioned studies have contributed to the body of knowledge and shed some lights on the topic of item analysis, they exclusively utilised test items obtained from university and high school level. Insufficient

amount of studies, to the best of the authors' knowledge, have been conducted on analysing multiple choice items on junior high school level, especially in Indonesian context. Thus, to address this gap, the present study aims to analyze multiple-choice items obtained from a junior high school test. This study focuses on analyzing the difficulty level, discrimination, power, and

distractor efficiency of the test. On the ground of this, three research questions are submitted to answer. They are: What is the difficulty level of the chosen multiple-choice items? What is the discriminating power of the chosen multiple-choice items? What is the distractor efficiency of the chosen multiple-choice items?

Research Method

The Nature of the Study

The present study employed a qualitative research design and document analysis as the research method. As put forward by Bowen (2009) document analysis refers to the acts of scrutinizing and interpreting documents to draw meanings and give voice to certain research topics. There are several documents commonly employed in document analysis: public records, personal documents, and physical evidence (O'Leary, 2014). As mentioned earlier, the present study seeks to analyse multiple-choice items, and in doing so, an official exam paper will be utilised. On this account, the authors believe that document analysis will be the best fit for the study.

Data Collection and Context of the Study

Multiple-choice items were obtained from a trial testing (also known as "tryout" in Indonesian context) at SMPN (state junior high school) 2 Tigaraksa were employed in the study. The test consisted of 50 questions focusing on vocabulary, grammar, and reading. The trial testing was performed in the 2018/2019 academic year. The time allocation provided to the students to finish the test was 120 minutes. 305 students participated in the test, and 100 students were selected as the sample of the study.

Data Analysis

Difficulty Level Analysis

Difficult level of test items is commonly defined as the portion of test-takers who choose the correct answer instead of the distractors on a test question (Fulcher & Davidson, 2007). Based on the achieved scores, question items on a language testing can be classified as difficult, moderately difficult, and easy. The table below presents more detail on this matter.

Table 1. The Difficulty Level

Index Level	Difficult Categories
-------------	----------------------

0.00-0.30	Difficult
0.31-0.70	Moderate
0.71-1.00	Easy

In calculating the level of difficulty, the present research will make use of the following formula:

$$\text{Difficulty level} = \frac{\text{numbers of students who answer the item correctly}}{\text{total number of students who answer the item}} \quad (1)$$

Discriminating Power

Discriminating power indicates the ability of a question item to set apart high achieving examinees (upper group) from their lower-achieving counterparts (lower group) (Zajda, 2006). The discriminating power levels fall into several categories; they are good excellent, satisfactory, and bad or rejected. The table below provides more detail.

Table 2. The level of discriminating power

Category	Discrimination Level
Excellent	0.71-1.00
Good	0.41-0.70
Satisfactory	0.21-0.40
Poor	0.00-0.20
Rejected	Negative

Following Gronlund (1998), the following is the formula the present study used to determine the chosen question items' discriminating power.

$$\text{Discriminating power} = \frac{RU - RL}{\frac{1}{2}T} \quad (2)$$

Explanation

RU: Numbers of students in the upper group who answer the question correctly

RL: Numbers of students in the lower group who answer the question correctly

½ T: One half of the total number of the students included in the item analysis

Distractor Efficiency

As previously discussed, distractors are alternative incorrect answers on a question item. A distractor is efficient if five or more percent test taker chooses it. On the other hand, if less than five percent of the examinee selects a distractor, the distractor is not efficient enough to distract students from choosing the right answer. More details on the classification of distractor efficiency are presented on the following table.

Table 3. The classification of distractor efficiency

Criteria	Category
$5\% \geq p \ \& \ LG > UG$	Effective
$5\% \geq p \ \& \ LG < UG$	Less Effective
$p \leq 5\% \ \& \ LG > UG$	Less Effective

Results & Discussion

Difficulty Level

The first analysis conducted on this study aims to discover whether the question items are

$p \leq 5\% \ \& \ LG < UG$ Ineffective

0 Dysfunctional

adapted from Arikunto (1986) and Malau-Aduli and Zimitat (2012)

Explanation

$5\% \geq p$: The number of students choosing the distractor is higher 5% or higher

$p \leq 5\%$: The number of students choosing the distractor is less than 5%

LG : The number of students in the lower group choosing the distractor

UG : The number of students in the upper group choosing the distractor

The formula used to compute the distractor efficiency is as follow:

Distractor efficiency:

$$\frac{\text{numbers of students who select the distractor}}{\text{total number of student who answer the item}} \tag{3}$$

perceived to be difficult, moderately difficult, and easy to answer by the study participants. The table below shows the final calculation of the difficulty level.

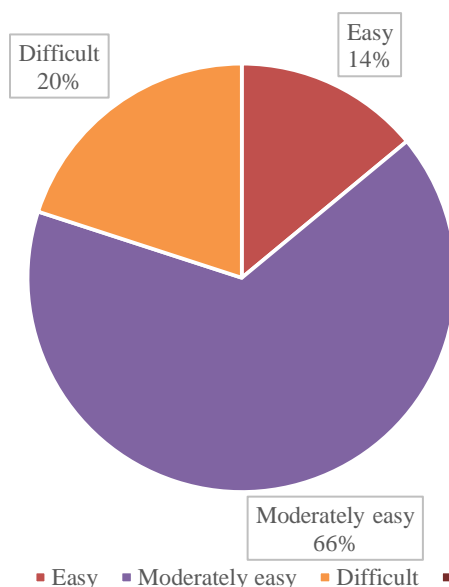
**Table 4
The final data of the difficulty level**

Difficulty level	Criteria	Frequency	Item numbers
0.71-1.00	Difficult	10	4, 11, 23, 27, 30, 33, 41, 45, 49, 50
0.31-0.70	Moderately Difficult	33	2, 3, 5, 6, 8, 10, 12, 14, 15, 16, 17, 18, 20, 21, 24, 25, 26, 28, 31, 32, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 46, 47, 48
0.00-0.30	Easy	7	1, 7, 9, 13, 19, 22, 29

As the above table displays, out of 50 questions, 7 questions are easy for the examinees to answer. In comparison, the other 33 questions are perceived to be moderately difficult, and the rest 10 questions are difficult for the test takers.

When translated into percentages, there will be 14% easy questions and 66% and 20% moderately difficult and difficult items, respectively. The following chart illustrates the percentage distribution.

Chart 1 Difficulty Level



It is evident from the data above that the levels of difficult on the question items are not evenly distributed. This outcome resonates with many previous item analysis studies on the difficulty level of multiple-choice questions. Abdul, Ayesha, and Syed Hammad (2018), for example, found in their study that difficult questions dominated the selected multiple-choice items, with more than 50% of the total number of the questions has a difficulty level more than 0.71. On the same token, Toksöz and Ertunç (2017) learned from their research that the chosen multiple-choice test used to evaluate preparatory school students consisted mostly of moderately difficult questions. Out of 50 questions obtained from the selected test, 23 were found to be relatively difficult and, as they argue, were suitable to test students' ability on the target knowledge. In Indonesian contexts, previous studies on item analysis like those of Danuwijaya (2018) and Yumelking (2019) exhibit similar results. While Danuwijaya discovered that moderately difficult questions dominated the test used in his study, Yumelking found out that the highest difficulty index proportion was filled with difficult questions.

The mentioned studies' results provide an obvious idea on how difficulty levels on test questions are never equally distributed. However, as one should note, such an action is not supposed to be perceived as an indication of an unfavorable practice in developing test questions. As Gronlund (1998) pointed out, the item analysis criteria should be intended to shed light on what specific language tasks students are or cannot

perform, not to discriminate between high and low-achieving students. What is crucial is to align the difficulty level with learning goals in the classroom. Thus, based of the study, the examination have good difficulty level because relatively moderate difficult 66% of all the question. A question item is not necessarily good when it has a high difficulty level or bad when it is easy. A question item that is answered correctly by all students, hence zero difficulty level, can be a good question when reflecting the outcome of the learning process. It simply informs us that all students have mastered the expected skills and can answer the question. This, however, is not supposed to discount the potential effectiveness of a difficult multiple-choice question. A difficult question that is formulated in accordance with classroom's goals and reflects learning purposes can also serve as as a good assessment tool to measure test takers' understanding and mastery towards target knowledge and materials. To conclude, what is important in constructing a test question is not whether the question is perceived as difficult or easy by test takers but whether it measures what is supposed to measure (Gronlund, 1998).

Discriminating Power

The second analysis performed in this study seeks to reveal the discriminating power of the chosen test items. The final calculation of the data is presented in the table 4 below.

Table 4. The final data of the discriminating power

Discriminating Power	Criteria	Frequency	Item numbers
0.71-1.00	Excellent	0	-
0.41-0.70	Good	16	6, 12, 14,15, 16, 17, 18, 20, 22, 31, 34, 35, 38, 41, 43, 47
0.21-0.40	Satisfactory	9	2, 5, 7, 24, 37, 39, 40, 44, 50
0.00-0.20	Poor	16	1, 3, 9, 11, 19, 21, 25, 27, 28, 29, 32, 33, 36, 45, 48, 49
Negative	Rejected	9	4, 8, 10, 13, 23, 26, 30, 42, 46

Table 4 informs that there is no question item with excellent discriminating power. However, there are 16 and 9 question items with good and satisfactory levels respectively. 16 questions with poor level of discrimination and 9 questions that should be rejected as they showed negative scores. The presented information tells us that half of the questions on the trial testing should be removed or at least revised and improved. The questions that fall into "rejected" criteria should be removed from the test as they very likely fail to reveal students' true understandings on the target subject while those included in the "poor" section should be revised or improved on their wordings. The presented information tells us that 9 questions on the trial testing should be revised and improved based on item numbers 4, 8, 10, 13, 23, 26, 30, 42, 46.

A lower and negative score on discrimination-powered is likely caused by ambiguous wording (Office of Educational Assessment, 2016). Ambiguous statement on a test is risky because not only can it prevent students from having a full understanding of what a question expects them to do, it can also potentially lead students to choose an incorrect option.

It has also been argued that multiple-choice questions with lower or negative discrimination scores fail to examine students' mastery of target knowledge and do not work properly to distinguish between students who possess a full understanding of certain classroom materials being tested and those who have a low-level mastery of knowledge. A further examination the maker question examination of a question with low and negative discriminating power is expected to determine the possible error or mistake on the question and improve them to get good discrimination power. Hence, it is suggested

that a revision on the wording of the questions with low and negative discriminating power should be performed to eliminate error or mistake on the questions and improve them.

In addition to examining test questions, revisiting and checking key answers are also suggested on multiple-choice items with low and negative discriminating power. A mis-keyed question has been a common phenomenon in developing test items. Mis-keyed questions occur when test developers put a wrong option or a distractor as the correct answer. When it happens, not only do the questions fail to validate test-takers' ability, but they also do not serve as an indication of learning goals.

The phenomenon of various level of discriminating power has also been reported on many item analysis studies in Indonesia. Hartanti and Yogi (2019) found in their study that the testing items they analyzed comprised of all levels of discriminating power with "satisfactory" and "poor" level dominates the question items. In the same vein, the study conducted by Danuwijaya (2018) also discovered the different level on difficulty on the multiple-choice items. Like the other study, Danuwijaya's study showed that the majority of the test items were dominated with "poor" and satisfactory" questions.

Distractor Efficiency

In addition to difficulty level and discriminating power, the present study performs an analysis of the distractor efficiency on the selected multiple-choice test items. For efficiency and the space of this paper, only 3 items will be presented in this discussion. The final data of distractor efficiency analysis on the 3 question items are displayed on the table below.

Table 5. The final data of the distractor efficiency

Group	Options				Key Answer
	A	B	C	D	
Upper	28	5	7	10	A
Lower	26	10	5	9	
Total	54 (54%)	15 (15%)	12 (12%)	19 (19%)	
Upper	19	12	17	2	A
Lower	22	5	11	12	
Total	41 (41%)	17 (17%)	28 (28%)	14 (%)	
Upper	10	5	30	5	C
Lower	8	12	12	18	
Total	18 (18%)	17 (17%)	42 (42%)	23 (23%)	

The efficiency of distractors on the chosen test items should be looked at from each question item. Question number 1, for example, has options B, C, and D as the distractors. From the above table, it is evident that there are more than 5 percent of test-takers who chose the options. The number of students in the lower group who chose option B is higher than those who chose the same option in the upper group. Thus, we can conclude that option B is an effective distractor. For options C and D, we can see that the number of students in the upper group is higher than those in the lower group. Therefore, we can conclude that option C and D are less effective as they distract more upper-class students. Following the same step, we can apprehend that B and C options on the question number are less-effective distractors, while option D is the most effective. As for question number 3, we can see that the option A attracts more students in the upper group, and that makes it a less effective distractor, while option B and D act as effective distractors for students in the lower group. The phenomenon of an ineffective distractor has been found in the study of item analysis in Indonesian context. Hartati and Yogi (2019), for example, discovered that the distractors of a question item on their study appeared to be completely unrelated to the question and it failed to attract test takers to choose them. Following Shin, Guo, and Gierl's (2019) suggestion, in the event that distractors do not act as they are expected to be, revising and generating distractor by using question prompt are needed. Writing distractors by using question prompt means that we create a false option by using the word that the question asks about or focuses on. If, for example, a question is asking about an appropriate response to a party invitation, then the distractor should also focus on party invitation.

Conclusion

The present study analyses question items obtained from a trial testing in a state junior high school. The study highlights three main findings. The level of difficulties witnessed on the test items varied. 10 questions were difficult while 33 and 7 other questions were respectively found to be moderately difficult and easy. The result shows that the majority of the questions on the test items were not too difficult or easy for the test takers to answer. The second finding shows that half of the chosen test items, 25 questions, have been effectively constructed which were reflected by their "good" and "satisfactory" level of discriminating power. These questions are arguably able to test students' expected ability. In contrast, the other half of test items, 25 questions, were found to be in "poor" and "rejected" levels of discriminating power. These question items were arguably to be vaguely worded and, thus, require some revisions. As many as 16 questions with "poor" level of discriminating power needed revisions on their wordings while the rest 9 question items should be removed. The study's last outcome reveals that in writing the distractors, the test developer tended to choose a statement that attracts more test-takers in the upper group than those in the lower one. As a result, compared with students in the lower group, more high achieving students selected wrong answers.

While the present study has some light on the quality of multiple-choice test questions in Indonesia's Examination, the study also contributed some knowledge on future examination. That is why future research on item analysis should include teachers' lesson plans. While the present study has shed some light on the quality of multiple-choice test questions and contributed some knowledge on language testing, the topic of the research could be extended on, for

example, the analysis on the alignment of the test items with learning purposes to reveal whether the test item has really met with the classroom's goals.

References

- Abdul, R., Ayesha, A., & Syed Hammad, H. (2018). Item analysis of multiple choice questions. *Pakistan Oral & Dental Journal*, 38(2), 291-293.
- Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, 16(1), 1-14. <https://doi.org/10.14434/josotl.v16i1.19106>
- Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan*. Bumi Perkasa.
- Assessment, O. E. (2016). *Understanding item analyses*. University of Washington.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bowen, G. . (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, 9(2), 27-40.
- Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Pearson Education.
- Crocker, M. (2019). *The importance of evaluation and testing in an educational system*. International TEFL and TESOL Training.
- Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education*, 7(1), 29-40. <https://doi.org/10.25134/erjee.v7i1.1493>. Received
- Dulger, M., & Deniz, H. (2017). Assessing the validity of multiple-choice questions in measuring fourth graders' ability to interpret graphs about motion and temperature. *International Journal of Environmental and Science Education*, 12(2), 177-193.
- Frost, R. (2005). *The question types*. British Council.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.
- Gronlund, F. (1998). *Assessment of student achievement* (6th ed.). Allyn & Bacon.
- Harris, D. P. (1969). *Testing English as a Second Language*. McGraw-Hill.
- Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF)*, 2(1), 59-70. <https://doi.org/10.24853/elif.2.1.59-70>
- 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>
- Javid, L. (2014). The comparison between Multiple-Choice (MC) and Multiple True-false (MTF) Test Formats in Iranian Intermediate EFL learners' vocabulary learning. *Procedia - Social and Behavioral Sciences*, 98, 784-788. <https://doi.org/10.1016/j.sbspro.2014.03.482>
- 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>
- Klufa, J. (2015). Multiple choice question tests- Advantages and disadvantages. *Mathematics and Computers in Sciences and Industry*, 91-94.
- Madsen, H. S. (1983). *Techniques in testing*. Oxford University Press.
- Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. In *Assessment and Evaluation in Higher Education* (Vol. 37, Issue 8, pp. 919-931). <https://doi.org/10.1080/02602938.2011.586991>
- Manalu, D., Sipayung, K. T., & Lestari, F. D. (2019). An analysis of students reading final examination by using item analysis program of eleventh grade of SMA Negeri 8 Medan. *JETAL: Journal of English Teaching & Applied Linguistics*, 1(1), 13-19.
- McNamara, T. (2000). *Language testing*. Oxford University Press.
- O'Leary, Z. (2014). *The essential guide to doing your research project* (2nd ed.). Sage Publications.
- Odukoya, J. A., Adekeye, O., Igbino, A. O., & Afolabi, A. (2017). Item analysis of university-wide multiple choice objective examinations : the experience of a Nigerian private university. *Quality & Quantity*, 1-15. <https://doi.org/10.1007/s11135-017-0499-2>
- Purpura, J. E. (2017). Assessing meaning. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (3rd ed., pp. 33-61). Springer.
- Shin, J., Guo, Q., & Gierl, M. J. (2019). *Multiple-Choice Item Distractor Development Using Topic Modeling Approaches*. 10(April), 1-14. <https://doi.org/10.3389/fpsyg.2019.00825>
- Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies*, 8(6), 141-146.
- Weimer, M. (2018). *Multiple-choice tests: Revisiting the pros and cons*. Faculty Focus.
- Yumelking, M. (2019). Test items analysis

constructed by EFL teachers of private senior high school in Kupang , Indonesia.
International Journal of English Literature and Social Sciences, 4(6), 1746-1752.

<https://doi.org/10.22161/ijels.46.19>
Zajda, J. (2006). *Learning and teaching*. James Nicholas Publisher Pty Ltd.