English Language in Focus
ELIF
e-ISSN: 2622-3597. Available online at https://jurnal.umj.ac.id/index.php/ELIF
Peer-review under responsibility of Muhammadiyah University of Jakarta, Indonesia

Item Analysis for a Better Quality Test

Neti Hartati[1)], Hendro Pratama Supra Yogi[2)]
[1), 2)] Universitas Muhammadiyah Prof. Dr. Hamka, Ciracas, Jakarta, Indonesia
[1)] *neti@uhamka.ac.id,* [2)] *hendropratamayogi94@gmail.com*

ABSTRACT

This study is a small-scale study of item analysis of a teacher's own-made summative test. It examines the quality of multiple-choice items in terms of the difficulty level, the discriminating power, and the effectiveness of distractors. The study employed a qualitative approach which also used a simple quantitative analysis to analyze the quality of the test items through the document analysis of the teacher's English summative test and the students' answer sheets. The result shows that the summative test has more easy items than difficult items with the ratio of 19:25:6 while they should be 1:2:1 for easy, medium, and difficult. In terms of the Discriminating Power, there are 3, 13, and 16 for excellent, Good, and satisfactory level, but there are 17 and 2 for poor and bad levels of Discriminating Power. There are 43 (21.5%) of all distractors which are dysfunctional which, in turns, makes the items too easy which also makes the items fail to discriminate the upper-group students from the lower ones. Therefore, the 43 dysfunctional distractors should be revised to alter the difficulty level and improve the discriminating power. This research is expected to serve as a reflective means for teachers to examine their own-made test to ensure the quality of their test items.

Keywords: Item analysis, summative test, the difficulty level, the discriminating power, the effectiveness of distractors
Citation APA Style: **Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus (ELIF)*, *2*(1), 59–70.**

INTRODUCTION

Evaluation holds a crucial role in education. It is a systematic process to gain insight into what a certain program does and how well the program runs (Patton, 1987). The Joint Committee on Standards for Educational Evaluation in Shinkfield & Stufflebeam (1995, p. 9) defined evaluation as "the systematic assessment of the worth or merit of an object". EDC (2013) explained that evaluation is a systematic process that involves collecting and analyzing data or information to make a decision or judgment about a specific program.

In the language classroom, teachers also evaluate to make a decision and judgment on their teaching program through assessment. There are various tools of assessment that teachers may use in the classroom, and one of them is by administering tests (Hughes, 2003, p. 5).

A test is a common instrument used by teachers to measure their

students' learning outcome. It is defined as "a method of measuring a person's ability, knowledge, or performance in a given domain" (Brown, 2003, p. 4). Brown further explained that test "is an instrument-a set of techniques, procedures, or items-that requires performance on the part of the test-taker." Popham (2003, p. 4) defined test as "an effort to determine the student's status in terms of their knowledge, skills, and attitudes."

Bachman & Palmer (1996, p. 8) explained that language tests provide valuable information on various aspects of a language teaching-learning process which may be used to evaluate the teaching-learning program itself. They, further, explain that tests may provide; evidence of the outcome of learning and teaching which may function as feedback of the effectiveness of the teaching program itself; information used to make decision of what kinds of learning materials and activities that should be given to students; a diagnosis of strengths and weaknesses used to decide whether an entire class or individual students are ready to move to another unit of instruction; assigning grades on the basis of learners' achievement; a way of clarifying the instructional objectives, instructional materials and activities based on the students' need of language learning.

Considering the crucial role of the test in the teaching process, teachers have to make sure that they are constructing a good quality test. Weir (2005) urged that teachers or test makers have to ensure that a test results scores which are an accurate reflection of an examinee's ability in a specific area. One of the ways of ensuring the quality of a test is by conducting an item analysis. Item analysis is a set of procedures in evaluating the quality of items that make up a test (Musial, Nieminen, Thomas, & Burke, 2009). Brown & Hudson (2002, p. 113) explain, "Item analysis is usually done for the purpose of selecting which items will remain on future revised and improved version of the test." To sum up, item analysis is a process of evaluating the quality of the test items done to sort out the good items from the weak ones and repair them to improve their quality for future use.

Item analysis is a process of examining students' responses to each test item done to measure the quality of the test items. It is a process of checking and analyzing the quality of each item by sorting out the good items from the weak ones and revised them to become better ones. Brown & Hudson (2002) and Musial, et al (2009) defined item analysis as a process done based on certain procedures and steps to identify which test items are effective and have good quality to be used as a tool of assessment.

Although the advantage of item analysis in constructing a good test has been widely realized, a large amount of research such as conducted by, just to name a few, Rafika (2014), Nihayatunnisa (2015), Farahdiba (2015), Setiowati (2015), Rafiqa (2015), Alittasari (2016), Syamsiah (2016), and Afziyatin (2018) found that Indonesian English teachers rarely or do not conduct item analysis to examine the quality of their summative test items. Therefore, this research intends to re-raise teachers' awareness of the importance of conducting item analysis to get an

understanding of the quality of teachers' summative test.

This research intends to conduct an item analysis of the English Summative test at a senior high school on the First Semester of 2017/2018 Academic Year. Based on the information, the English teachers at this school have never conducted an item analysis on their summative test and they have not known the quality of their summative test items. An item analysis of the English summative test, therefore, needs to be conducted at this school. The research aims to examine the quality of the summative test items for the first semester of the tenth-grade students.

A summative test is a test given at the end of a semester or a course administered to measure or sum up how much a student has learned from a course and achieved the learning objective (Brown, 2003; Cizek, 2010; Harmer, 2007; Hughes, 2003). The summative test is analyzed in this study is formulated in multiple-choice questions which according to (Harmer, 2007) "are extremely difficult to write well." Furthermore, teachers, generally, receive little or no training and support for assessment.

The analysis is conducted to get empirical evidence of the difficulty level, the discriminating power, and the effectiveness of distractors of the summative test items which are constructed in multiple-choice items. The empirical evidence will inform which items need to be accepted, revised, or rejected which then will be revised. The result of the analysis will also enable teachers to decide what teaching remedy which may be given to improve students' achievement of the learning objectives.

There are three formulated research questions for this study:

1. What are the characteristics of the summative test constructed by the teacher in terms of the difficulty level, the discriminating power, and the effectiveness of the distractors?
2. How many items should be revised, maintained or discarded based on the difficulty level, the discriminating power and the effectiveness of the distractors to improve the quality of the test items?

The procedures of conducting item analysis in this study involved three kinds of analysis; the analysis of the difficulty level, the discriminating power, and the effectiveness of distractors.

The analysis of the Difficulty Level or Facility Value (FV)

The analysis of the difficulty level or Facility Value (FV) is the first step in analyzing the test items. Heaton (1988) stated that the FV of an item shows the difficulty of an item in a test. It shows which item is easy or difficult. The FV can be known from the ratio or percentage of students who answer the item correctly. Fulcher & Davidson (2007), Reynolds & Livingston (2012), and Zajda (2006) defined FV as the proportion or percentage of test takers who correctly answered the question. This analysis will enable teachers to identify which items are easy, medium, or difficult. A good test should have a varied index of difficulty which consists of easy, moderate, or difficult. Sumarsono (2014) suggested that a good test should have a ratio of 1: 2: 1 for its easy, moderate and difficult items. It means that the test should have

25% easy, 50% moderate, and 25% difficult items.

This research uses Heaton's (1988) formula to measure the FV which is gained by dividing the number of students from the upper group and the lower group students who answer a certain item correctly by the total number of the students who join the test.

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

Explanation:

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

2n :The number of all students taking the test

The range of FV is from 0.00 to 1.00. To categorize the FV, the writer uses Sumarsono's (2014, p. 93) classification of the difficulty level:

Table 1: Categories of Item Difficulty

Difficulty Level	Category
0.00 – 0.20	Very difficult
0.21 – 0.40	Difficult
0.41 – 0.60	Moderate
0.61 – 0.80	Easy
0.81 – 1.00	Very Easy

The analysis of the discriminating power

The next step of item analysis is to determine the Discriminating Power (DP) that is whether the item can discriminate the students of the upper group from those in the lower group. (Zajda, 2006, p. 165) stated, "The discrimination power is whether the item differentiates test takers in higher achieving levels from those of lower achieving levels". Thus, DP is the extent to which the test item can distinguish students between the upper group and lower group students.

Heaton (1988, p. 180) elaborates, "The index of discrimination (D) tells us whether those students who performed well on the whole test tended to do well or badly on each item of the test." He, further, explained that the students' total score of the whole test is used as the valid measure or the criterion measure. The argument that underlies the index of discrimination is "If the 'good' students tend to do well on an item (as shown by many of them doing so, - a frequency measure) and the 'poor' students badly on the same item, then the item is a good one because it distinguishes the good from the bad in the same way as the total test score Heaton (1988, p. 180).

This research uses Heaton's (1988) formula of DP.

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

Explanation:

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

The result of the use of Heaton's formula above is interpreted by using Arikunto's (1986) criterion of DP.

Table 2: The Categories of Discriminating Power

Difficulty Level	Category
0.71 – 1.00	Excellent
0.41 – 0.70	Good
0.21 – 0.40	Satisfactory

0.00 – 0.20	Poor
negative	Bad/rejected

By analyzing and categorizing the scores of the DP, it can be decided which items should be accepted, revised, or rejected.

The analysis of the Effectiveness of Distractors

The analysis of distractors "provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item" (Professional Testing Inc, 2006). Good distractors contribute to the Discriminating Power of each item because they will attract the lower-group students to choose them and, hence, have a significant role to discriminate the upper group students from the lower ones, which in turns, contribute to the level of difficulty of each item and the quality of the test in general.

Reynolds & Livingston (2012) affirmed, "On multiple-choice items, the incorrect alternatives are referred to as distractors because they serve to "distract" examinees who do not actually know the correct response". It means the inaccurate choices or distractors give a contribution to the discriminating power of each item. Good distractors will attract the students who do not master the content of the learning material (Brown, 1996) and thus will attract more students from the lower group than the upper group (Gronlund, 1977). Therefore, good distractors affect the result of the test to differentiate the upper group students from the lower ones. The upper group will not be distracted to choose the distractors, while the lower group students will tend to choose them. Professional Testing Inc. (2006) stipulated the characteristics of good distractors. They stated that good distractors must be incorrect but plausible or seem likely reason for the students who are not sufficiently knowledgeable in the content area. They further explain that if a distractor appears impossible and does not attract any examinee to choose it, it will make the item far too easy than it should be, which in turns, makes the item have a poor level of Discriminating power. Therefore, when the distractors do not run their functions to distract the lower-group students or those who have not studied, the distractors should be revised.

Malau-Aduli & Zimitat (2012) claim that when distractors fail to attract examinees to choose them, it means the distractors are dysfunctional and do not give any contribution to the aim of the assessment. Further, Arikunto (1986)stated that a distractor is considered effective if it is chosen by at least 5% of test takers. The following is Arikunto's chart to determine the effectiveness of distractors added with Malau-Aduli & Zimitat's (2012) classification of the dysfunctional distractor.

Table 3: The Categories of the Effectiveness of Distractors

Standard	**Category**
5% ≥ p& LG > UG	Effective
5% ≥ p& LG < UG	Less Effective
p≤ 5% & LG > UG	Less Effective
p ≤ 5% & LG < UG	Ineffective
p=0	Dysfunctional

RESEARCH METHODOLOGY

The method used in this research is a qualitative method in the form of document analysis of the English summative test and students' answer sheets. The qualitative analysis is used to describe the quality of the items of the English Summative test. A simple quantitative analysis was used in analyzing the quantitative data of the facility value (FV), the discriminating power (DP), and the effectiveness of distractors.

This research was conducted at SMA Muhammadiyah 25 which is located on Jl. Surya Kencana No. 29, Pamulang barat, Tangerang, Banten. The respondents of the research were taken from two classes which consisted of 65 tenth-grade students of SMA Muhammadiyah 25 who were on their first semester. The data of the students' responses to the English summative test were taken from the answer sheets of those 65 students.

At the outset, the students' total scores were ranked starting from the highest until the lowest scores; ranked from 1 to 65. The 27% (18) of the highest score students are categorized as the upper group and 27% (18) students as the lower group. The students with the top 18 total scores, ranked from 1 to 18, belong to the upper-group students, while the students with the lowest scores, ranked from 48 to 65, belong to the lower-group students. The FV, the DP, and the effectiveness of distractors are then analyzed which, then, become the basis to make a decision of which items that can be kept for future use, revised or rejected.

FINDING AND DISCUSSION

The English summative test being analysed in this study consists of 50 multiple choice items which comprise of questions on; grammar (35 items or 70%), reading comprehension (11 items or 22%), and vocabulary (4 items or 8%). Each item comprises of 1 stem and 5 options of answers. Burton, Sudweeks, Merrill, & Wood (1991, p. 3) explains that the stem is the problem which may come in a form of a question or a complete sentence. The five options or alternatives consist of 1 answer key and 4 distractors.

The Difficulty Level

The use of Heaton's (1990) formula of FV reveals that the ratio of easy, medium, and difficult items of the English summative test has not reached the ideal ratio of 1:2:1. The current ratio is at 0.48:2:1.5. In other words, the percentages of the difficulty levels are 12%, 50%, 38% for the difficult, medium, and easy items respectively. It means that there are 6 easy items should be revised or changed into difficult items. The result of the analysis is summarised in the following table.

Table 4. The Percentages and Classification of the Difficulty Level of each item

No	Range of Difficulty Level	Criteria	Frequency	Item Numbers	Percentage
1	0.00 – 0.30	Difficult	6	1, 25, 45, 41, 47,48.	12%

2	0.31 – 0.70	Medium	25	3, 4, 9, 10,13, 14, 15, 16, 17, 18, 21, 24, 29, 31, 32, 35, 36, 37, 38, 40, 42, 44, 46, 49, 50.	50%
3	0.71 – 1.00	Easy	19	2, 5, 6, 7, 8, 11, 12, 19, 20, 22, 23, 26, 27, 28, 30, 33, 34, 39, 43	38%
	TOTAL		50		100%

The following is an example of the analysis of the FV of item #2.

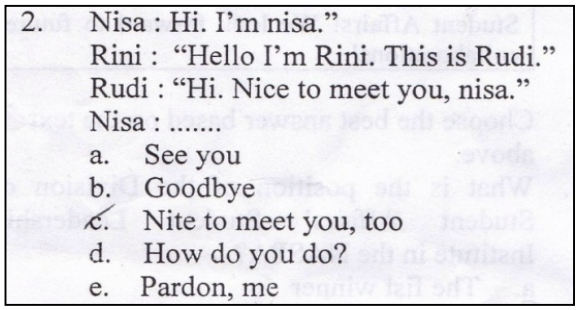

2. Nisa : Hi. I'm nisa."
Rini : "Hello I'm Rini. This is Rudi."
Rudi : "Hi. Nice to meet you, nisa."
Nisa : .....
a. See you
b. Goodbye
c. Nite to meet you, too
d. How do you do?
e. Pardon, me

Figure 1. Item #2

Table 5. The Facility Value of Item # 2

CU	CL	FV= CU+CL/2n	Category
18	17	0.97	Easy

The table shows that there are 35 students answer item #2 correctly, and the FV is 0.97 which indicates that this item is easy.

The Discriminating Power (DP)

The use of Heaton's formula of DP reveals that the 6% (3 items) with excellent discriminating power, 26% (13) items, and 32% (16) items with good and satisfactory levels of discriminating power respectively. In total, there are 64% (32) items with acceptable discriminating power and, hence, can be used for the future test. However, there are 16 and 2 items with poor and bad or rejected items respectively. This finding indicates that those 16 poor items should be revised and 2 bad items should be rejected or changed. The following table summarises the calculation of the DP based on Heaton's (1990) formula and Arikunto's (1986) criterion of DP.

Table 6. The Classification of the Levels of Discriminating Power of each item

NO	The Range of DP	Criteria	Item Numbers	Freq.	%
1	Negative	Bad	4,8	2	2%
2	0.00 – 0.20	Poor	2, 4, 5, 6, 11, 22, 23, 24, 25, 26, 34, 35, 41, 43, 44, 45, 49	17	34%
3	0.21 – 0.40	Satisfactory	1, 7, 10, 12, 19, 27, 28, 30, 32, 33, 36, 39, 40, 42, 50	16	32%
4	0.41 – 0.70	Good	8, 9, 13, 14, 15, 16, 17, 20, 21, 31, 37, 38, 46	13	26%
5	0.71 – 1.00	Excellent	3, 18, 29	3	6%
		Total		**50**	**100%**

Negative discriminators

Table 4 above shows that there are 2 items, item #4 and #8, with negative or rejected levels of discriminating power as shown in the table below.

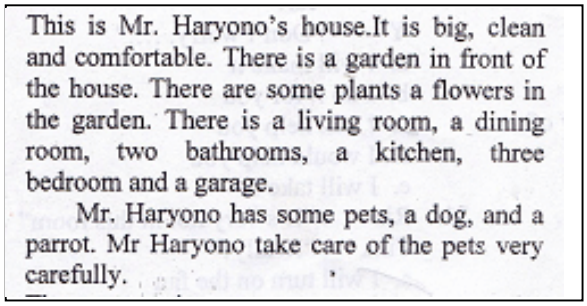
Table 7. The FV and the DP of items no. 41 and 48

No.	CU	CL	FV	Category of FV	D	Category of D
41	0	3	0.08	Difficult	-0.17	Rejected
48	3	4	0.19	Difficult	-0.05	Rejected

Professional Testing Inc (2006) urged that when there is an item with negative discriminator, we should recheck the answer keys for the items because there is a probability that it has a wrong answer key. They explained:

> ...if an item has discrimination below 0.0, it suggests a problem. When an item is discriminating negatively, overall the most knowledgeable examinees are getting the item wrong and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item is measuring something other than what the rest of the test is measuring. More often, it is a sign that the item has been mis-keyed (PTI, 2006; p.2).

It reveals that the two items have wrong or inappropriate answer keys. For item #41 below, the teacher decided that D is the answer key. However, the answer key should be C. This mistake has led to the wrong detections of the level of difficulty (0.08) and the discriminating power (0.18) of the item. Therefore, the answer key should be revised. This finding suggests that item analysis is also useful to identify mis-keyed items.


This is Mr. Haryono's house.It is big, clean and comfortable. There is a garden in front of the house. There are some plants a flowers in the garden. There is a living room, a dining room, two bathrooms, a kitchen, three bedroom and a garage.

Mr. Haryono has some pets, a dog, and a parrot. Mr Haryono take care of the pets very carefully.

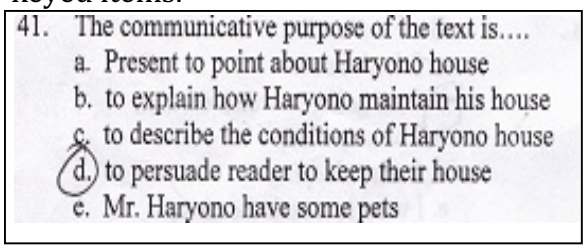

41. The communicative purpose of the text is....
a. Present to point about Haryono house
b. to explain how Haryono maintain his house
c. to describe the conditions of Haryono house
d. to persuade reader to keep their house
e. Mr. Haryono have some pets

Figure 2. Excerpt item #41

Item number 48 has the FV of 0.19 which indicates that the item is difficult and the Discriminating power is -0.05 which means that the item has a bad or rejected level of discrimination because there are more lower-group students (4) who answer this item correctly than the upper-group students (3).

It reveals that the answer key A "as happy as" is not the closest synonym as "as contented as". The closest synonym is "as satisfied as". Therefore, the answer key A must be changed. Professional Testing Inc (2006) urged that an answer key "must definitely correct". Further, this item has mistyped clues of expression because the expression "as contended as", which according to the stem should be written in italic type. This mistyped error may confuse the examinees.

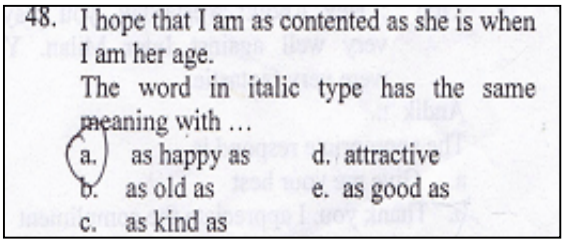

48. I hope that I am as contented as she is when I am her age.
The word in italic type has the same meaning with ...
a. as happy as
b. as old as
c. as kind as
d. attractive
e. as good as

Figure 3. The original version of item #48

The Effectiveness of Distractors

The analysis of distractors reveals that there are 107 (53%) effective distractors, 34 (17%) less effective distractors, 16 (8.5%) and 43 (21.5%) ineffective and dysfunctional distractors respectively. Effective and less effective

distractors could still be maintained. However, the 16 (8.5%) ineffective distractors must be revised. Finally, there are 43 (21.5%) categorized as dysfunctional distractors which fail to attract any examinee and, hence, should be changed because they do not give any contribution to the quality of the test in general.

Table 8. The Distribution of the Effectiveness of Distractors

No	Category	Value	%
1	Effective	107	53%
2	Less Effective	34	17%
3	Ineffective	16	8.5%
4	Dysfunctional	43	21.5%
Total		200	100%

The following is an example of the analysis of the effectiveness of distractors for item #2:

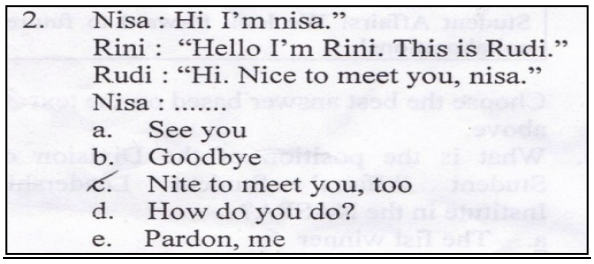

2. Nisa : Hi. I'm nisa."
Rini : "Hello I'm Rini. This is Rudi."
Rudi : "Hi. Nice to meet you, nisa."
Nisa : .....
a. See you
b. Goodbye
c. Nite to meet you, too
d. How do you do?
e. Pardon, me

Figure 4. Excerpt item #2

Table 9. The Effectiveness of Distractors of item #2

Category	LE	Df	Eff	Df	Df	Good	Σ
Option:	**A**	**B**	**C***	**D**	**E**	**O**	
Upper Group	0	0	18	0	0	0	18
Lower Group	1	0	17	0	0	0	18
Total	1	0	35	0	0	0	36

LE : Less effective
EFF : Effective
INEFF : Ineffective
* : The answer key
∑ : Sum of the students.

The table above indicates that item#2 has three dysfunctional distractors, options B, D, and E, which means that they must be changed or revised. Meanwhile, option A is less effective because there is only one student from each upper and lower group students who chose this distractor. A possible explanation for those three dysfunctional distractors is because they do not have the characteristics of a good distractor that they are incorrect but no plausible or seem unlikely and, hence, fail to attract any examinee to choose them

This finding also supports PTI's (2006; p. 2) claim that the presence of those three dysfunctional distractors make the item artificially far easier than it should be. This item has the difficulty level of 0.97 which means very easy and 0.05 level of discriminating power which means that this item cannot differentiate the upper group students from the lower ones. Therefore, to better the levels of difficulty and the discriminating power, the possible solution is to revise the distractors. The following is the alternative revision of distractors for item #2.

The following alternative revision for item #2:

Nisa : "Hi, I'm Nisa."
Rini : "Hello I'm Rini. This is Rudi."
Rudi: "Hi, Good to see you, Nisa."
Nisa:.....

a. See you
b. You too
c. Good to see you, too
d. You are good too
e. Good too

CONCLUSION

The results of the item analysis on the English summative test reveals that the proportion of the easy items is higher than expected (19 or 38%) while it should be only 12-13 items (12.5%). The level of the discriminating power of some items are also poor (17 or 34%), even 2 (2%) are rejected with negative Discriminating Power. Further analysis yields that the two aforementioned problems are due to the quite big number of bad or rejected distractors where there are 43 dysfunctional distractors which make those 43 items easier than they should be. Therefore, to improve the quality of the discriminating power and the level of difficulty, those 43 dysfunctional distractors should be revised.

Item analysis has proven effective in the effort of testing and improving the quality of multiple-choice items. By conducting item analysis, the quality of the stems, the answer keys, and also distractors can be tested, which in turns, may assist to prepare better test in the future to make sure that the assessment reaches its true goal.

REFERENCES

Afziyatin, P. (2018). *An Item Analysis of the English Summative Test for the Eleventh Grade Students of SMAN 90 Jakarta in the Second Semester of 2017/2018 Academic Year*. An unpublished paper. Universitas Muhammadiyah Prof. DR. Hamka.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. New York: Oxford University Press.

Brown, H. D. (2003). *Language Assessment: Principles and Classroom Practices.* New York: Pearson Longman.

Brown, J. D. (1996). *Testing In Language Programs: A Comprehensive Guide To English Language Assessment*. Upper Saddle River: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Retrieved from https://testing.byu.edu/handbooks/betteritems.pdf

Cizek, G. J. (2010). An Introduction to Formative Assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 3–17). New York: Routledge.

EDC. (2013). Understanding Evaluation: Promote Prevent- 3 Bold Steps. Retrieved from Education Development Centre, Inc. website: http://positiveschooldiscipline.promoteprevent.org/

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. New York: Routledge.

Gronlund, N. E. (1977). *Constructing Achievement Tests* (2nd ed.). https://doi.org/10.1002/tea.3660060216

Harmer, J. (2007). *The Practice of English Language Teaching*. England: Pearson Education Limited.

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

Foreign Language. London: Longman.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). UK: Cambridge University Press.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer Review Improves the Quality of MCQ Examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919–931. https://doi.org/10.1080/02602938.2011.586991

Musial, D., Nieminen, G., Thomas, J., & Burke, K. (2009). *Foundations of Meaningful Educational Assessment*. New York: McGraw-Hill Higher Education.

Patton, M. Q. (1987). *Qualitative Research & Evaluation Methods*. Thousand Oaks, CA: Sage Publications.

Popham, W. J. (2003). *Test Better, Teach Better: The Instructional Role of Assessment*. Virginia-USA: Association for Supervision and Curriculum Development.

Professional Testing Inc. (2006). *Step 9. Conduct the Item Analysis*. Retrieved from http://www.proftesting.com/test_topics/pdfs/steps_9.pdf

Reynolds, C. R., & Livingston, R. B. (2012). *Mastering Modern Psychological Testing: Theory and Methods*. Upper Saddle River, NJ: Pearson Education.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher Evaluation: Guide to Effective Practice*. London: Kluwer Academic Public.

Weir, C. J. (2005). *Language Testing and Validation An Evidence-based Approach* (C. N. Candlin & D. R. Hall, Eds.). New York: Palgrave Macmillan.

Zajda, J. (2006). *Learning and Teaching*. Australia: James Nicholas Publisher Pty Ltd.