THE ANALYSIS OF ENGLISH MID-TERM TEST ITEMS BASED ON THE CRITERIA OF A GOOD TEST AT THE FIRST SEMESTER OF THE EIGHTH GRADE STUDENTS OF MTS. MATHALIBUL HUDA MLONGGO IN THE ACADEMIC YEAR OF 2016/2017

Nihayatus Sa'adah
Islamic University of Nahdlatul Ulama` Jepara
Email: nsaadah21@yahoo.com

ABSTRACT

This study was aimed at finding out the items quality of English Mid-Term Test at the First Semester of the Eight Grade Students of MTs. Mathalibul Huda Mlonggo in the Academic Year of 2016/2017 Based on the Criteria of a Good Test. The researcher analyzed the multiple-choice items with total 25 items which focused at (1) the face validity, (2) the content validity, (3) the reliability, (4) the difficulty level and (5) the discrimination power. The research method used was descriptive-qualitative method. The subject of this study was eighth grade students of MTs. Mathalibul Huda Mlonggo in the Academic Year of 2016/2017 class VIII E which has 40 students. The researcher used documentary study to collect the data.

The finding showed that (1) the test has bad face validity since 92% or 23 (twenty three) items were error. There were only 8% or 2 (two) items categorized as appropriate items. (2) The test has good content validity since there were 88% or 22 (twenty two) items represent the basic competence and indicator as in the syllabus. (3) The test has high reliability since the reliability value of the items reached r11 = 0.691. Arikunto (2002:152) stated that reliability value which is between 0.61 – 0.80 has a high reliability. (4) The result of difficulty level analysis, the test was categorized into good test since 18 (eighteen) items (72%) were ideal items or having P around 0.62. There were 2 (two) items or 8% were founded as very easy items (P above 0.90). The rest 5 (five) items or 20% were categorized as very difficult items (P below 0.20). (5) The result of discrimination power analysis, the test has bad discrimination power since there were 20 (twenty) items (80%) were poor items category or having D 0.00 – 0.20 and 5 (five) items (20%) were satisfactory since they reached D 0.20 – 0.40. This means that the items needed to be revised.

Keywords: *Analysis, mid term test, good test criteria*

INTRODUCTION

Assessment is a common term in teaching and learning process. Teacher as an educator has a main role to assess the students' learning activities to know the teaching results by giving the students test to measure their understanding in the material and to know their improvement after certain material is

taught. Widoyoko (2014:1) stated that there are three terms related to assessment. They are test, measurement, and evaluation. People commonly define them as the same term, but beside they have different definitions. Test is a tool to gain information about the result of students' learning (Widoyoko, 2014:1). Guilford at Widoyoko (2014:1) defined about measurement is a process of giving / determining numbers to things according to a set of rules. Ralph Tyler (1950) in Arikunto (2005:3) explains that evaluation is a process of collecting the data of the attainment of education purposes. While assessment is an activity of describing the result of measuring something based on certain criteria. Clearly, among those four terms, evaluation has the widest range of all because it covers all the components of learning program. Widoyoko (2014:1) explained the components are inputting and processing learning result, students, teacher, curriculum, infrastructure, learning media, classroom, students' behavior etc.

Teachers evaluate their students to gain information whether the teaching and learning process is success or not. The teacher knows the strength and weakness of their teaching, and what need to be revised as well. Test is a tool which is used by teacher to evaluate the learning result. According to Brown (2004:3) test is a way to measure a person's ability, knowledge or performance in a certain domain. He explains more that a well-constructed test provides an accurate measure for the test takers' ability within a particular domain. This means that teachers as the test constructors should have ability to make a good test based on the students' ability and are able to analyze it. They should give a test based on the materials they have already given. The teachers' accuracy and carefulness in making the test give impact for the better quality of evaluation they make.

This study is aimed to find out the quality of Mid-Term test Items made by English teacher. The test is analyzed based on the criteria of a good test which is from five criteria 1) face validity, 2) content validity, 3) reliability, 4) difficulty level, and 5) discrimination power.

Indonesian Government by means of education ministry manages, coordinates, plans, and supervises education process for its development. Education evaluation is included in it. They arrange it in the laws of education and will be amended in a certain time based on the condition around. Based on section 1 verse (1), the Law of Education and Culture Ministry Number 53 year 2015 about assessment of learning outcomes by teachers and education units on elementary and secondary education,

"Penilaian Hasil Belajar oleh Pendidik adalah proses pengumpulan informasi/data tentang capaian pembelajaran peserta didik dalam aspek sikap,aspek pengetahuan, dan aspek keterampilan yang dilakukan secara terencana dan sistematis yang dilakukan untuk memantau proses, kemajuan belajar, dan perbaikan hasil belajar melalui penugasan dan evaluasi hasil belajar."(Permendikbud, Number 53 year 2015)

It can be concluded that assessment is one of important parts in teaching and learning. By learning assessment, teachers could get information/data on many aspects, such as attitude aspect, knowledge aspect and skills aspect. Moreover, the assessment is done systematically as planned and done to monitor the process, learning progress, and improvement of learning through the assignment and learning outcome evaluation.

There are many tests which are given by the teachers. There are daily task, weekly task, Mid-Term test, final test, and National

final examination. They are differentiated by the material, level and the time of conduction. This study discusses more about Mid-term test. Mid-Term test/Mid-term exam is an examination administered in the middle of an academic term, before semester test. It is conducted three months after the students get quarter material in one semester. Commonly, Mid-Term test conducted in many schools in Indonesia is made by the subject teachers. In some schools, rarely they use the task made by a group of subject teachers or education institution.

Most teachers of any subjects are rarely analyzing the test they have constructed. This is commonly because of limited skill owned by the teachers about doing the analysis. Even they do not know the way of analysis and what points to be analyzed. Moreover, the other reason of rarely doing analysis is having personal thinking that their work is enough in constructing and administering the test, then taking the result of the students after doing it. They do not care about the quality of the test, whether it is categorized as a good test or not. Moreover, experienced teachers tend to hunch the test they have constructed are good enough or even the best. If the teachers give attention to the test they constructed by doing analysis, they do not need to make it again in the next test period because the test has found the criteria of a good test.

C.McCowan (1999:3) stated that test item analysis is important to be done since the result can improve the item and test quality. He explained more that statistics and experts' judgments are used in item analysis to evaluate tests based on the individual items quality, entire sets of items, and the relationship of each item to other items. In addition, Thompson & Levitov (1985) as cited by (McCowan and McCowan, 1999:3) stated that item analysis examines the items performance as considered by individually both in some external relation criterion and in relation to the remaining items on the test (Thompson & Levitov, 1985:163). It can be said that analyzing item test is a complete evaluation for a test for a better test construction in the future.

Widoyoko (2014:130-131) mentioned some points of reasons why analyzing test item is necessary: (1) Finding out the strength and the weakness of the test items, determining the good items or determining which items should be revised. (2) Preparing complete information about test items' specifications to help teachers in constructing a test for any learning evaluation. (3) Finding out specific mistakes in the test, such as answer key error, difficulty level, and discrimination power of a test. After all, the teachers as a test constructor will be soon make a decision for the error test items. (2) A good analyzed-test can be useful for the test constructor. This way means the teacher can save the test which has been analyzed as a reference to the next test construction. Moreover, the good test items also can be used to test a group of students in the next period of time.

For a long time, there are no special laws and rules neither from education ministry nor from school's leader to do analysis of test items made by teachers. Some test item analysis data which available around, are test item analysis result conducted by researchers or experts who are wondering about the quality of the tests. Or only some teachers conduct the analysis and they can be counted.

Many experts characterize some characteristics of a good test in different criteria. According to Osman (2010:53), the characteristics of a good test are measured from the a) test objectivity, b) discrimination, c) comprehensiveness, d) validity, e)

reliability, f) specification of conditions of administering, g) direction of scoring and interpretation. Moreover, Gampper (2013:75) classified some criteria for testing a test are reliability, validity, authenticity, backwash, and practicality. Thereafter, OMERAD & Michigan State University Board of Trustees (2011:3) explained that qualities of all good tests are purposeful, valid, reliable, objective, comprehensive, differentiating, expected, instructive and useful. 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, what criteria they take.

METHOD

This research is designed as descriptive analysis research. Best (1982) on Sukardi (2003:157) defined that descriptive research is a research which tries to describe and interpret an object. Moreover, Nassaji (2015:129) added that the aim of descriptive research is to describe a characteristics of a phenomenon. The researcher then explained the analysis result in qualitative approach. Qualitative approach is describing, explaining and interpreting the collected data (Williams, 2007:67). Mubarok (2016:73) further explains that qualitative research is defined as naturalistic approach where the researcher does research as natural as possible, describing the found phenomenon during the research.

This study analyzed the English Mid-Term test items made by the English teacher of Eighth grade students of MTs. Mathalibul Huda Mlonggo in the Academic Year of 2016/2017. VIII E class was chosen as the research subject which has 40 students in total. The writer found out the Face Validity, Content Validity, Reliability, Difficulty Level and the Discrimination Power of the test. All of them were explained in qualitative approach to meet clear explanation data. According to Williams (2007:67), qualitative research tries to describe, explain, and interpret the collected data. This study designed using Non-statistical data which meant the result of this study was only analyzed by the related formula.

The writer used documentary study to collect the data. Documentary study is one of data collection technique by collecting and analyzing some documents, written (pictures) or from soft file data (Sukmadinata, 2013:221). The documents needed were the English Mid-Term test items, scored-students' worksheet, answer key, test construction guider and syllabus. Those data then analyzed to get the final result which is known as the quality of Mid-Term test items made by the English Teacher of MTs. Mathalibul Huda Mlonggo in the Academic Year of 2016/2017.

There are some steps in analyzing the data. (1) Reading and selecting the data which have been collected, (2) analyzing the face validity, (3) analyzing the content validity, (4) analyzing the reliability, (4) dividing the students' worksheets into 2 groups, high level and low level, (4) analyzing the difficulty level, and (5) analyzing the discrimination power.

The first criterion is face validity. Face validity is one of important criteria in testing validity. It is related to the test performance and layout, how it looks like from its outer part. Without doing a deep analysis of the face validity, it would be difficult to find the lack of the test items because the test seems good in the layout and ordering.

The face validity is analyzed from the stem, option, and the items' instructions. The error parts which the writer found were in the items' punctuations, letter or words typing error, grammatical rules, diction and miss typing the test's constructor made. This

analysis is presented in a table. The writer could find many parts of errors in most of items. There are 25 (twenty five) items in the form of multiple choices, the writer found 23 (twenty three) items were error. There were only 2 (two) or 8 % items were categorized as appropriate test items, which means 23 (twenty three) items or 92% others should be revised. The finding data concluded that the test items have bad face validity. Below are 2 (two) examples of error items found.

1. Mr. Doni : Excuse me, please.
   Vera : ... , What can I do for you, sir?
   Mr. Doni : Can you tell me the way to Kartini Hospital?
   Vera : Of course.
   a. Lool at me
   c. Yes, sir
   b. I don't think so
   d. Let's go

   The underlined points were the errors found by the writer. Here, the writer could find errors in 3 points. Those points were in the *stem* and *option* parts. Specifically, they are from *items' punctuations, letter or words typing error, and miss typing the test's constructor made.* The appropriate item was as the following:

   Correction:

   Mr. Doni : Excuse me, please.
   Vera : ... . What can I do for you, Sir?
   Mr. Doni : Can you tell me the way to Kartini Hospital?
   Vera : Of course.
   a. Look at me
   c. Yes, Sir
   b. I don't think so
   d. Let's go

   The dialog is for questions no 2 to 4. Mr. Rio : Pay attention, please! Let's start our lesson today. Open your homework now. We will discuss it together.
   Bima : Sorry to bother you, Sir. I forgot to bring my homework. May I submit tomorrow, sir?
   Mr. Rio : Okay, but don't do it again.
   Bima : Yes sir, thanks a lot.

2. Where does the dialog take place?
   a. Library
   c. Laboratory
   b. Classroom
   d. Canteen

   Still the same as the previous number, the underlined parts were the errors points. The writer found 4 errors points from the *stem* and *option* parts. Specifically, they are from *diction, items' punctuations, and grammatical rules*. The appropriate item was as the following:

   Correction:

   The dialog is for questions 2 to 4.
   Mr. Rio: Pay attention, please! Let's start our lesson today. Open your homework now. We will discuss it together.
   Bima : Sorry to bother you, Sir. I forgot to bring my homework. May I submit tomorrow, Sir?
   Mr. Rio : Okay, but don't do it again.
   Bima : Yes, Sir. Thanks a lot.

The second criterion is content validity. The writer correlated the way of analysis to the test's construction guider made by the teacher. The writer used the Basic Competence and Indicator as in the learning syllabus. Syllabus or course outline plays an important role in determining test items' content validity. According to Heaton (1998:27) as cited by Rosanti (2013:41) that a good content validity of a test is which covers all the content of syllabus. The content validity is analyzed using a table.

The final result of the analysis is completely described in the "comment" column part. The writer stated CORRECT and INCORRECT in the comment which meant whether the item is related to the Basic

Competence and the Indicator or not. From the table analysis, the writer found 22 items or 88% were correct, 2 items or 8% were incorrect and 1 item or 2% were not included in the test construction guider. Those "correct" items were 1, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25. Then the "incorrect" items were 5 and 10. The writer could not find item 2 in the test construction guider. The findings concluded that the test items has a good content validity since 88% items were covered all the Basic Competence and the Indicator.

Table 1. Analysis of Content Validity

Basic Competence	Indicator	Test item	Comment
3.1 Applying social function, text structure, and languange element of interpersonal interaction text, spoken and written which involve askingfor attention, checking understanding, showing appreciation for others, asking and giving opinion, and the responses as the contexts used.	1. Using simple sentences to ask for attention	1	Mr. Doni :**Excuse me, please**. Vera : **... , What can I do for you, sir?** Mr. Doni: Can you tell me the way to Kartini Hospital? Vera : Of course. a. Look at me b. I don't think so c. Yes, Sir d. Let's go The bolded sentences are expressions / simple sentences of asking for someone's attention and the response. So, the item 1 is CORRECT based on the Basic Competence and the Indicator.
	3. Creating sentence to appreciate someone's working	5	Which one that showing "Checking Understanding expression" below! a. I don't think so b. What do you think of it? c. Wow ! you are very beautiful dancer. d. Do you understand what I'm saying? Item 5 is INCORRECT. The indicator states about creating sentence to appreciate someone's working, but the stem is about checking understanding expression.

Basic Competence	Indicator	Test item	Comment
			Although it is related to the basic competence, but not related to the indicator.

#The third criterion analyzed is reliability. The writer used split-half method to analyze the items reliability. The strength of this method is only using one test and administered once (Arikunto, 2005:92), considering the items test which was analyzed here was administered once. The analysis of reliability is presented in a table.

Table 2. Analysis of the Reliability

Code	Odd Score (X)	Even Score(Y)	X^2	Y^2	XY
4	11	9	121	81	99
32	10	7	100	49	70
5	8	9	64	81	72
30	9	7	81	49	63
10	6	9	36	81	54
28	7	8	49	64	56
39	10	4	100	16	40
9	6	8	36	64	48
33	8	6	64	36	48
22	6	8	36	64	48
12	6	7	36	49	42
15	8	5	64	25	40
11	5	7	25	49	35
2	7	5	49	25	35
23	5	7	25	49	35
1	6	7	36	49	42
37	6	7	36	49	42
36	5	6	25	36	30
31	5	6	25	36	30
18	8	3	64	9	24
6	5	6	25	36	30
21	6	5	36	25	30
8	4	6	16	36	24
16	5	5	25	25	25
3	5	5	25	25	25
17	5	5	25	25	25
38	4	5	16	25	20
24	5	4	25	16	20

Code	Odd Score (X)	Even Score(Y)	X^2	Y^2	XY
34	5	4	25	16	20
13	4	5	16	25	20
26	5	4	25	16	20
19	4	5	16	25	20
27	5	4	25	16	20
7	3	5	9	25	15
40	5	2	25	4	10
35	4	2	16	4	8
25	4	2	16	4	8
29	4	2	16	4	8
20	1	4	1	16	4
14	0	0	0	0	0
Total	**225**	**215**	**1455**	**1329**	**1305**

Reliability is defined as the stability of a test which determined from the test result and the test is tested to the same level of students. The reliability analysis used split-half method then analyzed using reliability index as classified by Arikunto (2002:152). :

Reliability was found using *product moment* correlation and Spearman-Brown formula.

Calculating the *half test reliability* by using formula Product Moment

$$r_{XY} = \frac{(N.\sum xy) - (\sum x).(\sum y)}{\sqrt{(N.\sum x^2 - (\sum x)^2)(N.\sum y2 - (\sum y)^2)}}$$

$$\frac{(40.1305) - (225).(215)}{\sqrt{((40.1455 - (225)\ (40.1329 - (215))}}$$

$$\frac{(52200) - (48375)}{\sqrt{(58200 - (50625)(53160 - (46225)}}$$

$$\frac{3825}{\sqrt{(7575)(6935)}}$$

$$\frac{3825}{\sqrt{52532625}}$$

= 0.528

0.528 is the result of the half test reliability. For getting the reliability of the whole test, the writer uses formula *Spearman-Brown* and use 0.528 in the formula as the following:

$$r_{11} = \frac{2\, r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

$$\frac{2 \times 0.528}{(1 + 0.528)}$$

$$\frac{1.056}{1.528}$$

= 0.691

From the calculation reliability coefficient was found 0.691. This defined that the test items had high reliability. A test which has high reliability is categorized into a good test item. Furthermore, a good test can be used in the next time testing.

#The fourth criterion analyzed was difficulty level. It was started by dividing the students as the test's respondents into two groups. They were upper group (UG), the number of students who got good score and lower group (LG), the number of students who got lower score than the upper group. Those groups were determined by arranging the test scores from the upper to the lower. The researcher then analyzed them by using formula. The clear analysis of the difficulty level was presented in a table.

Table 3. Analysis of the Difficulty Level

No. Item Test	Ru	Rl	Ru + Rl	$\frac{Ru + Rl}{Nu + Nl}$	Comment
1	17	13	30	0,75	Ideal Value
2	19	7	26	0,65	Ideal Value
3	19	19	38	0,95	Very easy item
4	8	6	14	0,35	Ideal Value
5	9	6	15	0,375	Ideal Value
6	13	10	23	0,575	Ideal Value
7	7	1	8	0,2	Very difficult item
8	14	7	21	0,525	Ideal Value
9	14	9	23	0,575	Ideal Value
10	12	4	16	0,4	Ideal Value
11	10	7	17	0,425	Ideal Value
12	14	13	27	0,675	Ideal Value
13	10	4	14	0,35	Ideal Value
14	4	3	7	0,175	Very difficult item
15	4	2	6	0,15	Very difficult item
16	12	4	16	0,4	Ideal Value

No. Item Test	Ru	Rl	Ru + Rl	$\frac{Ru + Rl}{Nu + Nl}$	Comment
17	20	8	28	0,7	Ideal Value
18	20	17	37	0,925	Very easy item
19	11	4	15	0,375	Ideal Value
20	6	4	10	0,25	Ideal Value
21	9	1	10	0,25	Ideal Value
22	7	3	10	0,25	Ideal Value
23	1	1	2	0,05	Very difficult item
24	4	2	6	0,15	Very difficult item
25	11	8	19	0,475	Ideal Value

Difficulty level is the ration number of students who answered the question correctly to the total of students who participated in the test. A good test item is which not too easy and not too difficult. The difficulty level was analyzed using difficulty index classified by Sabri (2013:4). They are very easy item (*P* Above 0.90), ideal value (*P* around 0.62) and very difficult item (*P* below 0.20). The result showed that there were 18 items (72%) categorized as ideal items, very easy items were 2 items (8%) and 5 items (20%) were categorized as very difficult items. The finding concluded that the test items were categorized into good items because most of items were ideal items. This means that the items could be utilized as learning assessment.

#In analyzing the discrimination power of the test items, the researcher used the same steps as in analyzing the difficulty level. Firstly, the researcher arranged the test' scores from upper to the lower. Secondly, the researcher divided the test' score into two groups, upper and lower. Then, the researcher counted the discrimination power using formula D.

Table 4. Analysis of the Discrimination Power

No. Item test	UG	LG	(UG-LG)	$D = \frac{UG - LG}{n}$	Comment
1	17	13	4	0,1	poor
2	19	7	12	0,3	Satisfactory
3	19	19	0	0	Poor
4	8	6	2	0,05	Poor
5	9	6	3	0,075	Poor
6	13	10	3	0,075	Poor
7	7	1	6	0,15	Poor
8	14	7	7	0,175	Poor
9	14	9	5	0,125	poor
10	12	4	8	0,2	Satisfactory
11	10	7	3	0,075	Poor
12	14	13	1	0,025	Poor

No. Item test	UG	LG	(UG-LG)	$D = \frac{UG - LG}{n}$	Comment
13	10	4	6	0,15	Poor
14	4	3	1	0,025	Poor
15	4	2	2	0,05	Poor
16	12	4	8	0,2	Satisfactory
17	20	8	12	0,3	Satisfactory
18	20	17	3	0,075	Poor
19	11	4	7	0,175	Poor
20	6	4	2	0,05	Poor
21	9	1	8	0,2	Satisfactory
22	7	3	4	0,1	poor
23	1	1	0	0	Poor
24	4	2	2	0,05	Poor
25	11	8	3	0,075	Poor

Discrimination power is an ability of a test in differentiating the skill of low level students and the higher ones. The discrimination power was analyzed using discrimination power index classified by Arikunto (2005:218).

The result in the research finding showed there were 20 items (80%) were included into poor item category and 5 items (20%) were satisfactory items. Finally it could be concluded that the Mid-Term test items at the first semester of the eighth grade students of MTs. Mathalibul Huda Mlonggo in the academic year of 2016/2017 had bad discrimination power because most of the items were poor. Ebel's (1972) as cited by Sabri (2013:4) stated that poor items needed to be improved by revision

CONCLUSION

Test Item analysis is important to be done to improve the quality of the test items. Teacher who is as the test constructor needs to be able to do analysis to know the strength and the weakness of the test. There many five criteria in analyzing test items in this study.

First, the face validity of English Mid-Term test items had bad face validity since found 92% or 23 (twenty three) items were error. There were only 8% or 2 (two) items categorized as appropriate items. This means that the items need to be revised.

Second, the content validity analysis showed that the Mid-Term test items had a good content validity. The research found there were 88% or 22 (twenty two) items from 25 items in total, covered the basic competence and indicator as in syllabus.

Third, the reliability of the English Mid-Term test items had a high reliability since the reliability value of the items reached r_{11} = 0.691. High reliability criteria belong to good test items. This means that the test can be used in the other occasions to the same level of students.

Forth, difficulty level analysis found that the test items were categorized into bad items because the writer found 18 (eighteen) items (72%) were very easy items or having *P* above 0.90, 2 (two) items or 8% were founded as ideal items (*P* around 0.62). The rest 5 (five) items or 20% were categorized as very difficult items (*P* below 0.20). The ideal items

are when the most items found as ideal items which are not too easy and too difficult. This means that the test items must be revised.

Fifth, the discrimination power of the Mid-Term test items had bad discrimination power. There were 20 (twenty) items or (80%) were poor items category or having *D* 0.00 – 0.20 and 5 (five) items (20%) were satisfactory since they reached *D* 0.20 – 0.40. This means that the items needed to be revised.

REFERENCES

Arikunto, S. (2005). *Dasar-Dasar Evaluasi Pendidikan* (Edisi Revisi). Jakarta: Bumi Aksara.

Boopathiraj, C. and D. K. C. (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 ISSN 2277 3630 Vol.2 (2), February (2013), 2*(2), 189–193.

Brown, H. D. (2004). *Language Assessment principles and Classroom Practices*. United States of America: Pearson Education, Inc.

McCowan, Richard J and Sheila C. McCowan. (1999). *Item Analysis for Criterion-Referenced Tests*. New York: Center for Development of Human Services, 39. Retrieved from https://archive.org/details/ERIC_ED501716. Accessed on January 10th 2017.

Ciptaningrum, D. (2014). *An Item Analysis of English Summative Test on Difficulty Level and Discriminating Power (A Case Study on the First Grade Students of 3 State Junior High School of Tangerang Selatan)*. Jakarta: Syarif Hidayatullah State University Press.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment an Advance Resource Book*. London and New York: Routledge Taylor and Francis Group.

Gampper, C. (2013). *Improving English Test Qualities* (pp. 73–83). Thammasat Review, Special Issue. Retrieved from https://www.tci-thaijo.org/index.php/tureview/article/view/40732/33748. Accessed on January 10th 2017.

Harmer, J. (2007). *The Practice of English Language Teaching* (Fourth Edition). UK: Pearson Longman.

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

Nassaji, H. (2015). Qualitative and Descriptive Research: Data Type Versus Data Analysis. *Language Teaching Research, 19(2) 129*, 129–132. http://doi.org/10.1177/1362168815572747

Nurliyanto, D. (2015). *Test Item Analysis of the Final Examination on Economics Subject in Grade XII IPS SMA Negeri Banyumas Academin Year 2014/2015*. Yogyakarta: Yogyakarta State University Press.

OMERAD, M. S. U. B. of T. &. (2011). *Handbook of Learner Evaluation & Test Item Construction.* Michigan State University. Retrieved from http://omerad.msu.edu/meded/learner_evaluation/Learner_Evaluation_Handbook.pdf. Accessed on January 10th 2017

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

https://oer.avu.org/bitstream/handle/123456789/78/Educational%20Evaluation%20and%20Testing.pdf?sequence=1&isAllowed=y. Accessed on January 10th 2017.

Putri, N. S. (2015). *An Analysis of English Semester Test Items Based on the Criteria of A Good Test for the First Semester of the First Year of SMK Negeri 1 Gedong Tataan In 2012 / 2013 Academic Year*. Bandar Lampung: Lampung University Press.

Qoliliyah, N. (2013). *An Item Analysis of the English Final Semester for the twelfth Grade Students of the SMA Nasional Pati in Academic Year 2012/2013*. Kudus: Muria Kudus 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.

Shih, Y. (2010). An Item Analysis of an English Achievement Test Taken by EFL College Students in Taiwan. *6(3): 59-82(2010), 6*(3), 59–82.

Sukardi. (2003). *Metode Penelitian Pendidikan dan Pengembangannya* (First Edition). Jakarta: Bumi Aksara.

Sukmadinata, N. S. (2013). *Metode Penelitian Pendidikan* (kesembilan). Bandung: PT Remaja Rosdakarya

Sulfiana, F. (2014). *An Analysis of English Examination Items of the First Semester of the Eleventh Grade of SMA 1 Pecangaan Jepara in Academic Year 2012/2013*. Kudus: Muria Kudus University Press.

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

Williams, C. (2007). Research Methods. *Journal of Business & Economic Research*, *5*(3), 65– 72.

Penilaian Hasil Belajar oleh Pendidik dan Satuan Pendidikan pada Pendidikan Dasar dan Pendidikan Menengah (2015). Retrieved from https://luk.staff.ugm.ac.id/atur/bsnp/Permendikbud532015Penilaian%20HasilBelajarDikdasmen.pdf. Accessed on December 15th 2016.

Mubarok, Husni. (2016). English for Young learners Teachers Strategy in Teaching Reading. *Lensa: Kajian Kebahasaan, Kesusastraan, dan Budaya (Lensa), 6*(1), 68-83