AN ITEM ANALYSIS OF ENGLISH END-OF-TERM TEST WRITTEN FOR THE 9TH GRADE OF SMPN 28 SURABAYA

AN ITEM ANALYSIS OF ENGLISH END-OF-TERM TEST WRITTEN FOR THE 9TH GRADE OF SMPN 28 SURABAYA

Navisatul Izzah
English Education, Languages and Arts Faculty, State University of Surabaya
Navisatul.izzah@gmail.com

Dra. Theresia Kumalarini, M. Pd
English Education, Languages and Arts Faculty, State University of Surabaya

Abstrak

Tes memberi manfaat untuk guru maupun siswa. Dengan tes, guru dibantu untuk mengukur apakah tujuan pembelajaran telah tercapai. Untuk siswa, tes membuat mereka tau apakah guru mereka cukup baik dan konsisten dengan tujuan pembelajarannya ketika tes dikembalikan dan didiskusikan di kelas (Madsen, 1983:4). Sayangnya, siswa tidak mendapat manfaat tersebut karena faktanya tes jarang didiskusikan setelah pelaksanaan tes tersebut. Sebagian guru berpendapat bahwa tes sudah berakhir ketika siswa telah mendapat nilai tesnya (Heaton, 1988). Selain itu, guru kadang tidak membuat kisi-kisi soal yang merupakan hal penting dalam pembuatan soal tes. Hal ini membuat tes buatan guru tidak mempunyai karakteristik tes yang baik. Penelitian ini dilakukan untuk menganalisa salah satu tes buatan guru yaitu tes ujian akhir sekolah. Dengan menggunakan deskiptif sebagai desain penelitian, dan kuantitatif sebagai pendekatannya, penelitian ini bertujuan untuk menganalisa konten validitas, reliabilitas, tingkat kesulitan, dan tingkat diskriminasi dari tes tersebut. Melalui analisa oleh peneliti, didapatkan hasil bahwa tes ujian akhir sekolah relatif mempunyai konten validitas yang tinggi karena mencakup 75.4%. Namun, seharusnya tes dapat mencakup seluruh materi yang diajarkan guru. 24.6% dari soal yang tidak mempunyai konten validitas mengindikasikan tidak adanya kisi-kisi yang merupakan ilustrasi dari materi-materi yang ada di silabus. Tes ini juga mempunyai reliabilitas yang rata-rata karena koefisien reliabilitasnya 0.418, tingkat kesulitan yang rendah karena hanya sembilan dari tiga puluh lima soal yang berada di tingkat kesulitan yang tepat, dan tingkat diskriminasi yang kurang baik karena tes didominasi oleh soal yang kurang baik dalam mendiskriminasi siswa yang bisa dan yang kurang bisa.

Kata Kunci: *Analisis tes, Analisis butir soal, validitas, reliabilitas*

Abstract

A test gives advantages to both teachers and students. For teachers, it can help them measure whether the learning objectives have been achieved or not. For students, it makes them know whether the teacher is fair and consistent with the learning objectives or not when the test is returned and discussed in the class (Madsen, 1983:4). Unfortunately, the students do not get those advantages because the test is rarely discussed after being administered. Most teachers assume that the test is over after they got the scores (Heaton, 1988). Moreover, teachers sometimes do not make item indicators that are essential in making a test. It makes the teacher-made test has no characteristics of a good test. This study was conducted to analyze one of the teacher-made tests, i.e end-of-term test. Using descriptive research as the design and quantitative as the approach, this study aims to analyze the content validity, the reliability, the index of difficulty, and the index of discrimination of the test. From the analysis, the result showed that the end-of-term test has relatively high content validity because it contains 75.4%. Ideally, it should cover the whole materials taught. The missing 24.6% indicates the absence of item indicators which are a portrayal of the materials in the syllabus. It also has moderate reliability because the

coefficient of reliability is 0.418, low level index of difficulty because there are only nine out of thirty five items that are appropriate, and poor index of discrimination because the test is dominated by poor items which cannot discriminate the upper and lower group well.

Keywords: *Test Analysis, Item Analysis, Validity, Reliability*

INTRODUCTION

Students might have different feeling when they are going to have test. Some students feel worried, while some others feel at ease. For the first group, test is probably an unpleasant thing that might lead them to behave unsympathetically. While for the second group, test is unavoidable thing that should be dealt with well preparation (Suprihadi & Assyarofi, 2011). Yet, for both groups, test is actually followed by anxiety about the result/score, rank, and sometimes demands from parents and teachers. It can be proven by the high frequency of cheaters in every test. A research by Sulistiyanto et.al. (2008), showed that 93. 10 percent students are cheaters with various reasons. It is also supported by Johnson and Johnson (2002:27) that said there were more than 70 percent of students who declare as cheaters. It can be concluded that test is dreadful for almost all students.

Moreover, in foreign language testing, the challenges are enormous. Although the foreign language has been taught since in the fourth grade, in general, the students' competence is still low (Lie, 2007). Therefore, the test might give the students difficulty in comprehending the texts or even the questions because of some difficult words and also tenses that is not even similar with their first language. It sometimes contains of some materials that they even never get.

However, test is an essential and unavoidable thing in teaching learning process. Test is actually an instrument/tool to help teachers to find out whether the learning objectives have been achieved or not. It is supported by Madsen who stated that tests can help teachers to evaluate both theirselves –teachers and the students (1983:5). For students, test can be aid for them to make positive attitudes in the class; it makes them know that the teacher is fair and consistent with the learning objectives and to ease them to learn the language; when the test is returned and discussed in the class (Madsen, 1983:4).

Unfortunately, the students do not get those advantages. Test is rarely discussed after being administered. Most teachers assume that the test is over after they got the scores (Heaton, 1988). Furthermore, teacher just makes a test because it is an obligation. Sometimes they do not consider about their students' ability and the content of the test; they make too difficult test to challenge their students' knowledge/ to trap their students, they make too easy test to make the students get a good score, they just adopt the test from any sources without making item indicators that are a portrayal of the material that they have taught. Those facts show that teacher might not review the test before administer it. It is clear that not all tests can be considered as a good test or well-made test.

According to Madsen (1983:178), good tests should help the teacher to measure students' skills accurately. It will show that the teacher really concern about what she/he teaches. The common and famous things to be considered are validity and reliability. Validity means that the test is fair and relevant with the material that has been discussed in the class. Reliability deals with the consistency of results even on different occasion. Besides those two things, item analysis is also important to be considered as a factor to determine a good test.

Test can be divided according to the purposes, form, and test maker. According to the purpose of making test, test is divided into five types; placement test, progress test, achievement test, proficiency test, and diagnostic test (Alderson**,** et.al., 2005:11-12). According to the form of tests, test is divided into two; objective test and subjective test. According to the test maker, test is divided into two; standardized test and teacher-made test.

Standardized test is a standards-based test of a thorough process of many researches and development. It is used as reference and has a certain standards in administration and scoring (Brown and Abeywickrama, 2010:103). For instance, Ujian Nasional (UNAS) and Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN). Whereas a teacher-made test is prepared by teacher(s) to evaluate their students without being tried on first, analyzed and revised. For instance mid-term test (UTS) and end-of-term test (UAS). A mid-term test is usually made by a teacher that has responsibility to handle the class. It could be sure that the teacher will test what she/he has taught. While an end-of-term test is usually made by either chosen teachers in a district or a chosen teacher of the school. For the first and the second year students, the end-of-term test is made by chosen teachers in a district. While for the third year students, the end-of-term test is made by a chosen teacher of the school. It could be acceptable for some classes which the

teacher become the test-maker, but not for some other classes. It is supported by Nurgiyantoro (2001:61) that said a test that is made by a teacher should be applied in his/her own class rather than in other classes or even other schools that is not taught by the teacher. In fact, it is common that an end-of-term test often contains of some topic that unfamiliar for students in some classes.

It might make an end-of-term test doubtful, concerning to the validity, reliability, index of difficulty and index of discrimination of the test. Therefore, the writer wanted to conduct a study about the analysis of an end-of-term test for the third year or the ninth grade students in one of the schools that was chosen by simple random sampling.

This study aims to analyze one of teacher-made test that is an end-of-term test related to its validity, reliability, index of difficulty and index of discrimination. Moreover, this study only focused on thirty five items objective test of English end-of-term test. To limit this study focus, this study only concerned with the content of the test and the ninth grade students' works as the subject. Therefore, it did not concern with other levels of students.

Hopefully this study can be useful to give descriptions and knowledge about testing; the concept of testing, the importance in testing, the analysis of test.

RESEARCH METHOD

A descriptive research under the quantitative approach was conducted in this study. This study was conducted in SMPN 28 Surabaya. The school was chosen by simple random sampling of state junior high school of Rayon Surabaya Barat. In addition, at the time, the researcher found that the English teacher or other researchers did not conduct an analysis related to the test that the researcher studied on. In addition, according to the problems those are stated in the background of the study, it is appropriate to choose the ninth grade students' works as the subject of the study.

The researcher used cluster random sampling to draw the sample. As proposed by Kothari (2004:16), cluster sampling forms the population into some clusters and chooses cluster(s) rather than chooses per person to be the sample. The cluster was made according to the students' original classes; from IX A to IX H. Then, the researcher selected one cluster randomly by using lottery to be the sample. It was students' works class A.

There were five data that were needed to support this study. All of them were documents. The first data were the test items of the English end-of-term test of the ninth grade students. The second data were the syllabus of the first and the second semester for the ninth grade students that were used by the teacher. Those two data were used to analyze the content validity of the test. The third data were the students' raw scores that were used to analyze the reliability. The fourth data were the students' answer sheets. The last data was the answer's key of the test items. Those two data were used to calculate the index of difficulty and the index of discrimination of the test.

The researcher analyzed the data quantitatively. To analyze the content validity, the researcher matched the content of the test items with the teacher's indicators that are included in the syllabus and interpreted them to percentage. To analyze the reliability, the researcher used the students' raw score to find the mean of students' scores and standard deviation that are required by the formula of coefficient of reliability (r_{11}). Then, by using the formula, the reliability can be estimated. To analyze the index of difficulty and index of discrimination, the researcher used the students' answer sheets and the answer's key to identify each student's answer per item then applied formulas.

RESULT AND DISCUSSION

Content Validity

The content validity of the English end-of-term for the ninth grade students of SMPN 28 Surabaya was analyzed by matching the test content with the teacher's syllabus. To give the result of the content validity analysis, table 1 below is provided.

Table 1 Table of the result of content validity analysis for the first semester

	Basic Competence	Number of Items	Total	Percentage
First Semester	Listening	-	-	-
	Speaking	17, 18	2	5.8%
	Reading	-	-	-
	Writing	-	-	-
Total			2	5.8%

Table 2 Table of the result of content validity analysis for the second semester

	Basic Competence	Number of Items	Tot	Percentage

Second Semester			al	
	Listening	1,2	2	5.8%
	Speaking	-	-	-
	Reading	3, 4, 5, 6, 7, 8, 13, 14, 15, 16, 23, 24, 25, 26, 32, 33, 34	17	49.3%
	Writing	27, 28, 29, 30, 31	5	14.5%
Total			24	69.6%

The table presented that there are twenty six (75.4%) out of thirty five items that suitable with the teacher's indicators. Meanwhile the other nine items (24.6%) are more appropriate to the seventh and eighth grade students' material. It happened because the teacher did not make items indicators that are actually essential in making a test. Thus, the brief explanations about the twenty six items are explained below.

There are only two items that are appropriate to the indicators for listening skill of the second semester. They are items number 1 and 2. The percentage is 5.8%. These two items are related to one of the indicators of KD 7.2 which is to identify the information in the dialog about giving news and commenting on the news.

There are two items that are appropriate to the indicators for speaking skill of the first semester. They are items number 17 and 18. The percentage is 5.8%. These two items are related to one of the indicators of KD 3.2 which is to answer the question based on information in the dialog about asking for repetition.

There are seventeen items that are appropriate to the indicators for the reading skill of the second semester. They are items number 3, 4, 5, 6, 7, 8, 13, 14, 15, 16, 23, 24, 25, 26, 32, 33, and 34. The percentage is 49.3%. Item number 3 is related to one of the indicators of KD 11.2 which is to identify the information in the short functional text (letter). Item number 4 is related to one of the indicators of KD 11.2 which is to explain the content of the letter. Item number 5 is related to one of the indicators of KD 11.2 which is to explain the meaning of the word in the short functional text (letter). Item number 6 is related to one of the indicators of KD 11.3 which is to explain the social function of narrative text. Item number 7, 8, 13, 14, 15, and 16 are related to one of the indicators of KD 11.3 which is to determine the main idea or implicit and explicit information or word meaning or word reference in the narrative text. Item number 23 is related to one of the indicators of KD 11.3 which is to explain the social function of report text. Item number 24, 25, 26, 32, 33, and 34 are related to one of the indicators of KD 11.3 which is to determine the main idea or implicit and explicit information or word meaning or word reference in the report text.

There are five items that are appropriate to the indicators for the writing skill of the second semester. They are items number 27, 28, 29, 30, and 31. The percentage is 14.5%. Item number 27, 28, 29, and 30 are related to one of the indicators of KD 12.2 which is to determine the appropriate words to complete the blanks in the narrative text. Item number 31 is related to one of the indicators of KD 12.2 which is to arrange the sentences to make a good paragraph for narrative text.

From the explanation above, it can be concluded that the test covers the four skills; listening, speaking, reading, writing. Unfortunately, the proportion of each skill is not quite balanced. The test is dominated by reading skill that takes 49.3% out of 75.4%. The writing skill has the second biggest proportion that is 14.5%. While the listening and speaking have the same proportion that is 5.8%.

The researcher also found that the proportion of the materials is not balanced. There are thirteen KD in the first semester, but the two items that covers the first semester materials are from one KD (KD 3.2). There are thirteen KD in the second semester, but only four KD that are used: two items are from the same KD (KD 7.2), three items are from the same KD (KD 11.2), fourteen items are from the same KD (KD 11.3), and five items are from the same KD (KD 12.2).

However, the test covers the materials for quite a big percentage (75.4%). It is concluded that the test has high content validity. As supported by Bloom (1981:73), a test has high content validity if it covers 75% (or more) of the materials, low content validity if it covers a lesser amount of 50% of the materials, and moderate content validity if it covers 50%-70% of the materials. In this study, the English end-of-term of test for the ninth grade students of SMPN 28 Surabaya covers 75.4% of the materials. Therefore, it can be concluded that the test is considered to be claimed as a test that has high content validity.

Reliability

To estimate the reliability of the English end-of-term test for the ninth grade students of SMPN 28 Surabaya, the researcher chose internal consistency as the method. By using internal consistency, the researcher needed to obtain the students' scores then apply a formula.

To apply the formula, the researcher needed to find out some components that were needed. The first component was the mean of the students' score. To find out the mean of the students' score, the researcher needed to multiple the students' raw score (x) and the frequency (f) then divide them by the total students that took the test (n).

The mean score of the test is 24. There are 26 students who get score 24 or more than the mean score. While the number of students whose scores are smaller than 24 are 10.

The second component was standard deviation of all students' scores (s.d.). To find out the standard deviation, the researcher needed to find the deviation (d) first. To ease the calculation, the researcher arranged the students' score from the highest to the lowest to deviate the scores by mean score. After that, the deviation (d) needed to be squared per score (d^2) in order to fulfill the requirements of the formula to calculate the standard deviation (s.d.). The total of the squared deviation (Σd^2) is 455. To find out the standard deviation, the squared deviation (Σd^2) needed to be divided by the total of students that took the test (n) then applied a square root.

The standard deviation (s.d.) of the students' scores is 3.55. Next, the researcher calculated the coefficient of reliability (r_{11}) in order to find out whether the test was considered to have reliability or not. The calculation is presented below.

$$r_{11} = \frac{N}{N-1}\left[1 - \frac{m\,(N-m)}{Nx^2}\right]$$
$$r_{11} = \frac{35}{35-1}\left[1 - \frac{24\,(35-24)}{35(3.55)^2}\right]$$
$$r_{11} = \frac{35}{34}\left[1 - \frac{24\,(11)}{35(12.63)}\right]$$
$$r_{11} = 1.02\left[1 - \frac{264}{442}\right]$$
$$r^{11} = 1.02\,(1 - 0.59)$$
$$r^{11} = 1.02\,(0.41)$$
$$r^{11} = 0.418$$

The coefficient of reliability of the end-of-term test for the second semester of ninth grade students of SMPN 28 Surabaya 2014/2015 is 0.418.

The range of reliability coefficient is from 0 (zero) to 1 (one). But, the maximum number, 'one' does never exist; there will be no test that absolutely perfect without error. So does the minimum number, 'zero' number; there will be no test that entirely error (Douglas, 2009:107).

From the explanation above, it can be concluded that 0.418 is not either great or bad. In addition, Fulcher (2010:83) supported with the statement that there are five degrees that are used to interpret reliability coefficient; 0.01- 0.20 = very low, 0.21- 0.40 = low, 0.41- 0.60 = moderate, 0.61- 0.79 = high, 0.80- 0.99 = very high. Because 0.418 is between 0.41- 0.60, it is clear that it belongs to moderate reliability.

Index of Difficulty

To find out the index of difficulty, the researcher first organized the students to be either upper group or lower group according to their scores in equal size. Then, the researcher identified each student's answer per item then applied formulas.

$$F.V = \frac{R}{N}$$

The results of the calculation determine in which the items belong to the criteria of index of difficulty (Very difficult, difficult, moderate, easy, and very easy). The range of index of difficulty is from 0 to 1. The closer the index difficulty to 0, the more difficult the question for the test takers. The closer the index difficulty to 1, the easier the question for the test takers (Boopathiraj & Chellamani, 2013:190). Commonly, the question will be considered too easy when more of 90% or 0.90 of the test takers get it right. The question will be considered too difficult when a lesser amount of 30% or 0.30 of the test takers get it right (Madsen, 1983:181-182). In addition, Heaton (1988:179) stated that the accepted items or moderate items are those which have the index of difficulty between 30% (0.30) and 70% (0.70). According to the theory above, it can be concluded that if the index of difficulty is between 0.91 and 1.00, the question is considered as too easy or very easy. If the index of difficulty is between 0.71 and 0.90, the question is considered as easy. If the index of difficulty is between 0.30-0.70, the question is considered as moderate or accepted. If the index difficulty is between 0.21-0.29, the question is considered as difficult. If the index of difficulty is between 0.00-0.20, the question is considered as too difficult or very difficult. Thus, the result of index of difficulty is explained below.

Table 3 Table of the result of index of difficulty analysis

Criteria of index of difficulty	Number of item	Total
Very difficult (0.00-0.20)	6, 11, 29	3
Difficult (0.21-0.29)	32	1
Moderate (0.30-0.70)	**4, 9, 17, 20, 23, 25, 30, 33, 34**	**9**
Easy (0.71-0.90)	1, 5, 8, 12, 14, 15, 19, 22, 24, 27	10
Very easy (0.91-1.00)	2, 3, 7, 10, 13, 16, 18, 21, 26, 28, 31, 35	12

Based on the table, it is clear that from thirty five questions, only nine questions that are in the appropriate level of difficulty for the ninth grade students of SMPN 28 Surabaya as the test takers. They are number 4, 9, 17, 20, 23, 25, 30, 33, and 34 which belong to moderate level that has index of difficulty between 0.30-0.70.

There is only one difficult item that is item number 32. It is considered as difficult item because the index of difficulty is between 0.21-0.29. There are three items that are considered to be very difficult items. They are items number 6, 11 and 29 which belong to the index of difficulty between 0.00-0.20. There are ten items that are considered as easy items because they have index of difficulty between 0.71-0.90. They are items number 1, 5, 8, 12, 14, 15, 19, 22, 24, and 27. There are twelve items that are considered to be very easy items because they have index of difficulty between 0.91-1.00. They are items number 2, 3, 7, 10, 13, 16, 18, 21, 26, 28, 31, and 35.

The end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya 2014/2015 can be considered as an easy test for the students because the test is dominated by very easy (12 items) and easy items (10 items).

Madsen (1983:182) stated that the test should be in the appropriate level; i.e not too difficult or too easy. From the result, it can be concluded that there are nine items which are in moderate level. It means that there are only nine items that are in the appropriate level for the test takers. Moreover, more than half items of the test in the easy and very easy level. Therefore, most of the test items should be revised. It is supported by Madsen (1983:182) that said the test items that are too difficult or too easy should be rewritten.

Index of Discrimination

To find out the index of discrimination, the researcher first organized the students to be either upper group or lower group according to their scores in equal size. Then, the researcher identified each student's answer per item then applied formulas.

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

The results of the calculation determine in which the items belong to the criteria of index of discrimination. The brief explanation is explained below.

Table 4 Table of the result of index of discrimination analysis

Criteria of D value	Number of Item	Total
Excellent (>0.39)	**4,25,30,**	**3**
Good (0.30-0.39)	12,23	2
Mediocre (0.20-0.29)	1,5,8,9,15,22,27	7
Poor (0.00-0.20)	2,3,7,11,13,14,16, 17,18,19,21,24,26 ,28,31,35	16
Worst (< -0.01)	6,10,20,29,32,33, 34	7

Based on the table, it is clear that the English end-of-term test for the second semester of ninth grade students of SMPN 28 Surabaya 2014/2015 contains only three excellent items. They are items number 4, 25, and 30 which have index of discrimination >0.39. There are two items that belong to good items because they have index of discrimination between 0.30-0.39. They are items number 12, and 23. There are seven items that belong to mediocre items because they have index of discrimination between 0.20-0.29. They are items number 1, 5, 8, 9, 15, 22, and 27. There are sixteen items that belong to poor items because they have index of discrimination between 0.00-0.20. They are items number 2, 3, 7, 11, 13, 14, 16, 17, 18, 19, 21, 24, 26, 28, 31, and 35. There are seven items that belong to worst items because they have index of discrimination < - 0.01. They are items number 6, 10, 20, 29, 32, 33, and 34.

From the result, it can be concluded that the end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya 2014/2015 has poor index of discrimination because most of the items are

poor and even worst to discriminate the upper and lower group.

In addition, Ebel & Frisbie (1986) suggested that items that are worst should be absolutely discarded, items that are poor should be reviewed in depth or even discarded, items that are good can be either maintained or improved, and the excellent can be maintained. Unfortunately, in this test, there are only three items that belong to excellent items so that they can be maintained.

CONCLUSION

Based on the result and discussion in the previous explanation, the researcher concluded the four conclusions that are related to the research questions. First, the English end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya has high content validity because the test covered 75.4% from the materials; 5.8% from the first semester, and 69.6% from the second semester. The test covered the four language skills (listening, speaking, reading, and writing) although the proportion was not quite balanced. Moreover, there are nine items in the test which are more appropriate with the materials of the seventh and eighth grade students than ninth grade students. Second, the English end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya has moderate reliability because the reliability coefficient is 0.418. Third, the English end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya has low level for index of difficulty because there are only nine out of thirty five items that are in appropriate level for students. Fourth, the English end-of-term test for the second semester of the ninth grade students of SMPN 28 Surabaya has poor index of discrimination because there are only three items that are excellent to discriminate the upper and lower group.

REFERENCES

Alderson, J. C., Clapham, C., and Wall, D. 2005. *Language Test Construction and Evaluation.* UK: Cambridge University Press.

Bloom, B. S., George, M., & Thomas, H. J. 1981. *Evaluation to Improve Learning*. New York: Mc Graw-Hill, Inc.

Boopathiraj, C., & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. *International Journal of Social Science and Interdisciplinary Research*, 189-193.

Brown, H. D., and Abeywickrama, P. 2010. *Language Assessment: Principles and Classroom Practices (2nd ed.).* USA: Pearson Education, Inc.

Douglas, Dan. 2009. *Understanding Language Testing*. USA: Routledge.

Ebel, R. L., & Frisbie, D. A. 1986. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Fulcher, Glenn. 2010. *Practical Language Testing*. USA: Oxford University Press.

Heaton, J. B. 1988. *Writing English Language Tests*. United States of America: Longman Inc

Johnson, D. W., and Johnson, R. T. 2002. *Meaningful Assessment (A Manageable and Cooperative Process).* United States of America: A Pearson Education Company.

Kothari, C.R. 2004. *Research Methodology (Methods and Techniques) (2nd revised ed.).* New Delhi: New Age International, Ltd.

Lie, Anita. (2007) Education Policy and EFL Curriculum in Indonesia. *TEFLIN Journal*, Vol 18, No 1, pp. 1-14.

Madsen, H. S. 1983. *Techniques in Testing*. England: Oxford University Press

Nurgiyantoro, Burhan. 2001. *Penilaian Dalam Pengajaran Bahasa dan Sastra (Edisi ketiga)*. Yogyakarta: BPFE-Yogyakarta.

Sulistiyanto, A., Azkiyah, N. and Julianto, E. (2008) Upaya Preventif Edukatif Membentuk Generasi Anti Korupsi. *Paper presented on Diskusi Pelajar dan Mahasiswa Se-Kudus, Tim Program Kreativitas Mahasiswa Bidang Pengabdian pada Masyarakat 2008, Universitas Muria Kudus.*

Suprihadi and Assyarofi, T. B. (2011) Test-Taking Strategicness in Open Book Tests. *TEFLIN Journal*, Vol 22, No 2, pp. 167-184.