A Content – Form Analysis of English Final Test Items for The Second Semester of The Eleventh Grade Students of SMA Negeri

Wahyu Arta S, Abdul Asib, and Dewi Sri Wahyuni

English Education Department
Teacher Training and Education Faculty
Sebelas Maret University of Surakarta

Email: arta.wahyu41@gmail.com

Abstract: The objective of this study is to identify the quality of the test items used as a final test in the second semester for the eleventh grade students in SMA N in Magetan. This research used descriptive method. In collecting the data the writer used document (English final test items, syllabus, and students‟ answer sheets) as data sources. The data were analyzed by using the formula given by Ahmann and Glock. The results of this study shows that 57.5% of the total items have a good level in discriminating index, 45% of the items have fulfilled satisfactory criteria in difficulty level, 11 items had possessed the effective distracter, while the item‟s indicator 92.5 % of the items are compatible with the learning indicator mentioned in the syllabus, and in the construction aspect, 75% of the total items possess a good stem and 82.5% of the total items are able to fulfill all the aspects of good alternatives. In short, the items used as final test have good quality in constructing aspects, and its compatibility with the syllabus. However, some items are less effective viewed from its level of difficulty and the effectiveness of the disctracter aspect.

Keywords : *Item analysis, level of difficulty, discriminating index*

INTRODUCTION

Evaluation is important elements in the curriculum model beside initial planning procedures (consisting of data collection and learner grouping), content selection and gradation, methodology (including the selection of materials and learning activities), and ongoing monitoring, and assessment. Following the thought of Shaw and Dowsett in Nunan (1997:7) evaluation is the final component in the curriculum model. Traditionally, evaluation occurs at the final stage in the curriculum development process. Evaluation gives information about how successful the efforts of education have been. Evaluation can not be separated from teaching-learning process. Generally, the aim of second language evaluation is to improve the teaching and enhance the learning process. It helps teacher to get the information about the progress of students‟ achievement of the material they have learned in order to make decisions.

The prevailing of "*Kurikulum Tingkat Satuan Pendidikan*" (KTSP) gives a freedom for teachers in teaching learning activity, teachers design the lesson plan and implement it in the classroom by her/him self, including constructing the test item as an evaluation tool. In teaching learning activity there are several

components namely; teaching-learning process, teaching-learning purpose, and teaching-learning evaluation. They can not be separated. Evaluation is done to get the information about students‟ achievement related to the material they have learned.

Based on Ngalimun Purwanto (2006:3) an educational evaluation is the estimation of the development and progress of pupils toward objectives or values in the curriculum. The aim of evaluation is to gain data or information that shows the level of ability and the success of students in achieving the curricular objectives.

Many techniques are available to collect information for evaluation purposes, one of them is test. Tests are considered as a useful means for gathering information about student‟s achievement but it cannot be applied for gathering any of other types of information. Evaluation method that will be the focus in this study is evaluation with test, due to the fact that tests are considered the most familiar way used in evaluating student‟s learning.

Test and assessment also can be used to fulfill various aims such as monitoring students‟ achievement, providing school or system accountability, reporting to parents, or making decisions about individual students e.g., grade-to-grade promotion or school graduation.

According to Linn and Gronlund (2000:31) assessment is a broad term which covers the full range of procedures applied to obtain information about students‟ learning (observation, ratings of performances or projects, paper-and-pencil tests) and the value judgments formation focusing the progress of learning. A test is a specific type of assessment which typically consists of a set of questions conducted during a fixed period of time under fairly comparable conditions for all students.

Test can be of great help in gathering information for teaching evaluation of English as a second or foreign language. Tests, however, are relatively limited because they can only reveal certain aspects of student achievement; they cannot show us much about the other factors that often appear in the second or foreign language evaluation.

Testing has, traditionally, measured the results of student performance. Testing involves some steps: choosing the representative samples of language; measuring whether a student can use these samples; quantifying this by turning it into a mark or grade; keeping a record of these marks and use this to give an end assessment.

In Indonesian education system, a test seems to be the basis of the decision making of student graduation such as *UAS (Ujian Akhir Semester)* or *UN (Ujian Nasional).* Unfortunately, protests against the use of test as the criterion of graduation aroused when some known smart students failed. They protested that the test is an unfair way to make an important decision such as the student graduation.

On the other hand, some students gave no complaint toward the use of the test. They still considered the test as an objective way and there was nothing wrong about the test.

Various kinds of test items are applied in measuring students‟ performance. Students seem to be familiar to objective test items e.g., true-false, completion, matching, short answer, and multiple-choice items. Another familiar type is essay test items which give them chance for answering questions in their own word or ideas.

A test is supposed to be well-constructed so that it can be used effectively. To be said a good test, it has to fulfill the characteristics of good test, they are validity, reliability and practically. It is valid if the test can measure what is supposed to be measured. It can be reliable if the result from the test is the same even though the test is administered to the same standard for several times. A test can be practical if it is easy to do and administer.

The teacher should prepare the test as good as possible before administering the test to the examinees. The teacher needs to analyze the items of the test in order to know the effectiveness of the test and how well the item works. An analyzing the test items can be begun from matching the test items‟ indicator with the syllabus. By seeing the formed test item indicator, it will give information about the item which is analyzed whether in constructing the test item is in line with the syllabus or there are several test items are not, so that it makes the test item less quality.

There are two ways to analyze the test items. They are analyzing the test items by using qualitative technique relating to the content and form, and quantitative technique relating to the statistical feature (Anastasi and Urbina, 1997:172). Qualitative analysis covers the field of content and constructs‟ validity of the test items while quantitative analysis covers the measurement of level difficulty and discrimination power of the test items including the validity and reliability of the test items.

The principal purpose of analyzing the test items which were constructed by the teacher is to identify the weaknesses of the test items in testing or teaching-learning process. Based on the statement above, doing analysis to test item has many advantages, those are: (1) it can help the tester in evaluating the test that will be used, (2) it is very appropriate on arranging the informal test that is prepared by the teacher for student in the classroom, (3) It supports the constructing of the effective test items, (4) It can materially revise the testing in the classroom, (5) it can improve the quality of the validity and reliability of the test items.

The final test items in the end of the semester which were arranged by the teacher may not be said perfect yet because most teacher do not analyze the test items before they are used so that the quality of the test items is unknown yet. Based on the thought above the writer is interested to carry out a study to analyze and investigate the quality of the test items which is used in one senior high school in Magetan in the 2013/2014 academic year.

RESEARCH METHOD

In this study the researcher used descriptive method in the collection and analysis of the data. There are several kinds of descriptive study, which one of them is documentary analysis which often refers to content analysis. Johnson and Christensen (2000:302) determine that descriptive research is a research which focused on providing an accurate data in the form of description or picture of the status on characteristics of situation or phenomenon.

The focus of this study is the appropriateness of the test items which are used as the final test in the second semester for the eleventh grade students in one senior high school in Magetan in the 2013/2014 academic year.

The data used in this research are English final test items for second

semester of eleventh grade in one senior high school in Magetan in the 2013/2014 academic year and its answer key, standard competences, basic competences, learning indicators of English for grade XI in one senior high school in Magetan, blueprint of the test items and students‟ answer sheet.

100 students‟ answer sheets of the eleventh grade are taken as the sample by using stratified random sampling because the population of this research consists of 2 majors, namely natural science and social science class. The sample taken is believed to be representative. Arikunto (2006:134) states if the total number of the population is less than 100, it is better for the researcher to take the whole member of the population in his research. While the member of population is more than 100, the sample can be taken around 10%-15%, 20%-25% or more from the total member of the population.

The concept of Item analysis will be used in the analysis. Ahmann and Glock (1971:184) state reexamining each test item to discover its strengths and flaws is known as item analysis. Item analysis usually concentrates on two vital features: level of difficulty and discriminating power. The former means the percentage of pupils who answer correctly each test item; the latter the ability of the test item to differentiate between pupils who have done well and those who have done poorly. This method determines the difficulty of the item in a much more objective manner. It can be calculated by dividing the number of pupils who answer the item correctly with the total number of pupils who tried to answer the item then multiplied by 100%. The result of the calculation will be the level difficulty of the item. The resuls of the calculation is symbolized with (P).

Thorndike and Hagen in Anas Sudjono (2005:372) determine the difficulty index in three ranges i.e. When P is less than 30%, it means the item is very difficult, while the value of P is between 30% and 70%, the item is considered as a satisafactory item. The last, when P value is more than 70%, the item is considered as an easy item.

The second calculation is to find the discrimination index of each item. Item discrimination refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested. The aim is to specify the characteristics of the upper and lower group. We can use an independent criterion such as score from standardized test considered to measure the same achievement aspects or from a final mark in the similar achievement areas. On the other hand, an internal criterion may be used, such as the total scores from the classroom achievement test.

Before computing the discriminating power, we need to classify the students into three groups i.e., lower, middle, and upper.Item discriminating power of a test is its ability to separate good students from poor students. These students groups are defined by their scores on the test as whole. The difference between the percentage of the top scoring (upper group) 27% and bottom (lower group) scoring 27% of students who get the item right in its discrimination index while the middle group 46% is discarded.

Item discriminating power (D) can be obtained by subtracting the number of students in the lower group who get the item right (L) from the number of students in the upper group who get the item right

(U) and dividing by the number of pupils in each of two groups (N).

The maximum size of the index is +1.00 and the minimum size is -1.00. Any negative value means that the test item discriminates – to some degree in the wrong direction and is not satisfactory. Positive values show that the test item discriminates in the desired direction, even though it may not be complete satisfactory. The larger the positive value the better. Any D values above +0.40 can be considered very good, any between +0.40 and +0.20 satisfactory, and any between +0.20 and zero poor.

Meanwhile, in analyzing the construction of stem and alternatives of the items, Rana and Noor (2011:33) determine several aspects to be considered when constructing the stem as follows: (1) the stem clearly formulate a problem, (2) the stem is written in direct question form or in an incomplete statement form, (3) the stem only presents one problem, (4) the stem includes as much of the item as possible, without including irrelevant material, (5) unnecessary words or phrases have been avoided in the stem, (6) the stem is stated in positive form, and (7) the stem and should use proper grammar, punctuation and spelling.

To construct good alternatives some aspects are also given as follows: (1) The alternatives are worded clearly and concisely, (2) The alternatives are homogeneous in content, (3) the alternatives are free from clues as to which response is correct, (3) the choices in each item should be of approximately the same length or paired by length, (4) “all of the above” and “none of the above” have been avoided in the alternatives, (5) the item includes one and only one correct or clearly best answer, (6) the alternative should use proper grammar, punctuation and spelling.

Each item is analyzed based on the aspects shown above, if the items are not appropriate, it should be revised or discarded.

RESEARCH FINDINGS AND DISCUSSION

This chapter presents the global findings of item analysis on the multiple choices, content appropriateness of items with the syllabus, and the construction of the stem and alternatives of each item which is used in the test.

According to the analysis on the multiple-choice items, the writer finds some items which are unable to fulfill the discriminating power and difficulty level aspects properly.

From the discriminating power aspect, 57.5% of the total items have a good level in discriminating index. It is divided into 2 categories, namely 37.5% of the total items are satisfactory (15 items) and 20% have very good index of discriminating power (8 items).The rest items (42.5%) are poor in discriminating level (17 items).

The items which have a satisfactory index of discrimination power are represented by the items number 1, 2, 8, 10, 12, 15, 18, 21, 27, 28, 30, 33, 36, 37, and 38. Those items mentioned above are having the level of discriminating index around 0.20 – 0.39.

The items which have a good index of discrimination power are represented by the items number 4, 7, 14, 16, 32, 34, and 39. Those items are having the level of difficulty index more than 0.40. The items mentioned above are discriminating the students in the upper and lower group very well.

The rest items which are not mentioned above meaning that the items (item no. 3, 5, 6, 9, 13, 17, 19, 20, 22, 24, 25, 26, 29, 31, 35 and 40) are unable to discriminate the students in the proper manner. Those items are considered as a bad item viewed from its discrimination value as they only reach the value of discrimination index less than 0.20.

However, the item number 35 discriminates in negative value (D= -0.11) which means the item works in the wrong direction. More students in the lower group got the correct answer than the students in the upper group.

From the difficulty level aspect, only eighteen of the total items fulfilled satisfactory criteria in difficulty level. It means that 40 % of the total items have satisfactory level, 2 items (5%) are very difficult and the rest 55% are easy items which consist of 21 items.

The items which have a satisfactory index of difficulty level are represented by the items number 2, 4, 7, 8, 11, 12, 14, 15, 16, 21, 27, 28, 30, 32, 34, 37 and 39. Those items mentioned above are having the level of difficulty index around 30% - 70%.

While the items which have a very difficult level are represented by the item number 35 and 38. Those two items have the value of difficulty level less than 30% which means that item number 35 and 38 are only answered correctly by not more than 30% from the total students who take the test.

The rest 21 items are considered as an easy item which is represented by the item number 1, 3, 5, 6, 9, 10, 13, 17, 18, 19, 20, 22, 23, 24, 25, 26, 29, 31, 33, 36 and 40. Those items reach the value of difficulty level more than 70% which means those items are answered correctly by more than 70% from the total students who take the test.

From the effectiveness of the distracters aspect, there are 11 items which possessed the effective distracter (around 3-4 distracters work), 14 items are less effective because they only have 1-2 distracters work. The rest 15 items are poor with no distracter works.

While from the analysis on the content appropriateness of items with the syllabus is found out that The 92.50% or 37 out of 40 items used in the test are valid items because the items are able to cover the learning indicator which is mentioned in the syllabus, while the rest 7.50% (3 items) item no. 4, 7, and 8 are invalid since the content of those items are not in line with the syllabus. Those items should be discarded.

The last is the result of analysis on constructing the stem and alternative of multiple choice items. According to the analysis on the construction of stem and alternatives of each item, the writer found some items which are unable to fulfill the aspects of good stem and alternatives. 25% (10 items) of the total items are unable to fulfill the aspects of good stem i.e. item no 1, 2, 14, 24, 31, 33, 37, 38, 39, and 40. While the result on analysis the alternatives, the writer found that 9 items (22.5%) did not have good alternative as they are failed to fulfill all the aspects of good alternatives. Those items are item no. 1, 2, 3, 4, 11, 12, 13, 15, and 16.

Item analysis is helpful for the teacher especially for informal achievement tests or non-standardized tests which is constructed by the teacher. From the analysis, improvement of the items can be done since the analysis shows the weaknesses and strengths of the test items. Teacher can find students‟ achievement,

students‟ difficulty in mastering a certain subject or topic.

Teacher can also consider which item is needed to be improved, discarded, or saved for the next tests. Item analysis helps teacher to improve the multiple choice items since this type of test items are objective, easily quantified, and calculated using a certain formula.

CONCLUSION AND SUGGESTION

Based on the result of the analysis in this research, not all English final test items fulfill the criteria of a good test item perfectly. Only 25% of the total items fulfill all criteria meaning only 10 items have a good index of discriminating power and level of difficulty with effective distracters. While the rest of the items i.e., 30 items (75%) still have some weaknesses, even seventeen of them do not fulfill any criteria of a good test item.

The analysis on the content appropriateness of the items with syllabus found out the result that 92.5% (37 items) of the total items are in line with learning indicators which are stated in the syllabus, while the rest 3 items are unable to cover the learning indicator as well.

The quality of the test items seen from the construction aspect shows the result that ten items (25%) are unable to fulfill the aspects of good stem. While 9 items (22.5%) did not have good alternative as they are failed to fulfill all the aspects of good alternatives.

Through this article, the researcher would like to suggest that the test maker should give more attention in constructing the test items. Test publishers should construct test items which are in line with what students have actually learned and is based on the syllabus. Test publishers should construct better test item especially if it is used as standardized achievement tests which is conducted in a large scale. Test publishers should have some research or conduct a try out before issuing the test items used for testing the students.

REFERENCES

Ahmann,J.S, & Glock,M.D. 1971. *Evaluating student progress: Principles of tests and measurements, 6th ed.* Boston : Allyn and Bacon.

Anastasi, Anne and Urbina, Susana. 1997. Psicohololological Testing. (Seventh Edition). New Jersey: Prentice-Hall, Inc.Depdinas.2009. *Panduan Analisis Butir Soal*. Jakarta : Departemen Pendidikan Nasional

Arikunto, Suharsimi. 2006. *Dasar-dasar Evaluasi Pendidikan Edisi Revisi*. Jakarta: PT Bumi Aksara

Johnson, Burke and Christensen, Larry. 2000. *Educational Research : Quantitative and Qualitative Approaches*. Allyn & Bacon : A Pearson Education Company.

Linn, Robert L. and Norman E. Gronlund. 2000. *Measurement and Assessment In Teaching 8th Edition*. New Jersey: Prentice-Hall, Inc.

Nunan, David. 1997. *The Learner-Centred Curriculum.* Cambridge: Cambridge University Press.

Purwanto, M. Ngalimun. 2006. *Prinsip-Prinsip dan Teknik Evaluasi Pengajaran.*Bandung: Penerbit PT Remaja Rosdakarya.

Saed, Rana and Noor, Murtaza. 2011 .*Test Item Construction,Technique.* Pakistan :NTS press

Sudijono, Anas. 2005. *Pengantar Evaluasi Pendidikan*. Jakarta: PT Raja Grafindo

Persada.