AN ANALYSIS OF ENGLISH FINAL TEST FOR THE FIRST SEMESTER OF THE ELEVENTH GRADE STUDENTS OF SMA NEGERI 3 JOMBANG 2012/2013

Elsa Fatonah Triwulandari

S-1 English Education, Languages and Arts Faculty, Surabaya State University, ssaelsa@yahoo.com

Abstrak

Pengajaran dan tes adalah dua hal yang tidak bisa dipisahkan satu sama lain (Heaton, 1975:1). Sebuah tes memiliki sebuah tujuan yang sejalan dengan pengajaran yang telah dilakukan sebelumnya. Untuk memenuhi tujuan tersebut, tes harus memiliki karakteristik tes yang baik. Validitas dan reliabilitas adalah dua karakter utama dalam penafsiran dan penggunaan kemampuan bahasa (Bachman, 1990:24). Selain itu, untuk panduan dan perbaikan butir soal tes dibutuhkan sebuah analisis butir soal (Shakil, 2008) yang terdiri dari tingkat kesukaran dan daya beda soal. Oleh karena itu, peneliti melakukan sebuah analisis soal ujian akhir semester pertama mata pelajaran Bahasa Inggris untuk kelas XI SMAN 3 Jombang 2012/2013. Penelitian ini dilakukan untuk mengetahui validitas isi, reliabilitas, tingat kesukaran, dan daya beda soal. Peneliti menggunakan metode deskriptif quantitatif. Data penelitian terdiri dari soal, SI Bahasa Inggris kurikulum 2006, daftar nilai siswa, lembar jawaban siswa, dan kunci jawaban. Hasil penelitian menunjukkan bahwa soal ujian akhir semester pertama mata pelajaran Bahasa Inggris kelas XI SMAN 3 Jombang 2012/2013 memiliki validitas isi dengan kriteria layak dengan prosentase 70% dan reliabilitas 0.633 dengan kriteria layak. Untuk tingkat kesukaran, tes menunjukkan bahwa hampir separuh jumlah butir soal termasuk dalam kriteria mudah dengan prosentase 48%. Selain itu, untuk daya beda soal, sebagian besar butir soal termasuk kriteria tidak layak dengan prosentase 40%.

Kata Kunci: validitas isi, reliabilitas, tingkat kesukaran butir soal, daya beda butir soal.

Abstact

Teaching and testing are so closely interrelated each other (Heaton, 1975:1). A test has a purpose which is based on the teaching before. To fulfill the purpose of the test, it should have the characteristics of good test. Validity and reliability are the two essentials to the interpretation and the use of language ability (Bachman, 1990:24). Besides, one powerful technique available to the teachers for the guidance and improvement of instruction is the test item analysis (Shakil, 2008) which consists of item difficulty and item discrimination. Based on the explanation, the researcher conducted an analysis of English final test for the first semester of eleventh grade students of SMAN 3 Jombang 2012/2013. The purpose of this study is to know the content validity, reliability, the index of difficulty and discrimination. The researcher used quantitative method within descriptive approach to describe the quality of the test based on the result of the analysis. The source of the data in this study is the English final test items; the Standar Isi of English in curriculum 2006 for the eleventh grade students for the first semester; the students' scores and answer sheets, and the answers' key. The result of the study leads the researcher to some conclusions that the English final test for the first semester of the eleventh grade students of SMAN 3 Jombang 2012/2013 has moderate content validity with the percentage of 70% and moderate reliability with the coefficient of 0.633. For the index of difficulty, the test shows 48% of the test items are considered as easy items. Besides, for the index of discrimination, most of the test items are poor items with the percentage of 40% because they cannot differentiate between the students who got higher and lower scores.

Keywords: content validity, reliability, index of difficulty, index of discrimination.

INTRODUCTION

Teaching and testing are so closely interrelated. Both of them are practically impossible to work in either field without being constantly concerned with the other (Heaton, 1975:1). Test is mainly as a device to reinforce learning and to motivate the students, or to assess the students' performance in the language (Heaton, 1975). The test will be able to fulfill the purpose of the test if the test has the characteristics of good test. Validity and reliability are the two essentials to the interpretation and the use of language ability (Bachman, 1990:24).

Validity of a test is the extent to which it measures what it is supposed to measure and nothing else (Heaton, 1975, p. 153). It means that a test can be valid when the test actually can test what the teachers want to test. Test validity presupposes that the teachers can be precise about what to be tested and take steps to ensure that the test reflects realistic use of the particular ability to be measured (Weir, 1993, p. 19).

There are four kinds of validity of the test. They are face validity, content validity, construct validity, and empirical validity (Heaton, 1975:153-154). While there are several types of validity, the most important type for most programs is probably that of content validity.

Content validity is the extent to which a test measures the representative sample of the subject matter content (Heaton, 1975: 154). In fulfilling the content validity, the test should represent the material that the teacher is going to test. A test will be valid if the test consists of a representative sample of the course, and the relationship between the test items and the course objectives always being apparent. It means that content validity implies that the test should cover the materials which are stated in the curriculum or SK/KD. A test is considered as a test which has a high content validity when the agreement of the test is 75% or more. On the other hand, a test will have low content validity when the test has the agreement less than 50% (Bloom, George, & Thomas, 1981:73).

Reliability is the consistency of a test (Brown, 2004, p. 20). It means that a test can be reliable if the test is consistently measuring the performance of the students from time to time. There are some factors affecting the reliability of a test which are showed by Heaton (1975:155-156), such as a) the extent of the sample of material selected for testing, b) the administration of the test, c) test instruction, d) personal factors, e) scoring the test. The reliability of the whole test in objective tests can be estimated by using the formula:

$$\mathbf{r} = \frac{\mathbf{N}}{\mathbf{N-1}}\left[\mathbf{1} - \frac{\mathbf{m(N-m)}}{\mathbf{Nx^2}}\right]$$

r : the reliability

N: the number of items in the test

m: the mean score on the test for all the students

x: the standard deviation of all the students' scores

The highest reliability is 1.00. The level of reliability itself is between 0 and 1.00 ($0 < r > 1$). The teacher-made tests are usually considered has adequate reliability if the reliability of the test is 0.60 or above.

Many teachers think that the test is finished after the scores of the students have been obtained (Heaton, 1975) so that the teachers do not reflect on their purpose of giving the test to the students before. Heaton (1975) stated that the results which are obtained from the test actually can be used to provide further valuable information concerning on (1) the performance of the students as a group, thus informing the teacher about the effectiveness of his teaching; (2) the performance of individual students; and (3) the performance of each of the items comprising the test. The information is very important for teaching purposes, not only showing the types of errors which the students mostly made but also the actual reasons why the errors are made. Unfortunately, in fact, most teachers or test construction instructors will tend to use them again without further changes or just adapt them for future test because constructing good test items is time and effort consuming. One powerful technique available to the teachers for the guidance and improvement of instruction is the test item analysis (Shakil, 2008).

Item analysis examines how the test items perform as a set (Matlock-Hetzel, 1997). It is mainly important in improving items and reducing confusing items if the teacher will use again in later tests. Actually, it is helpful for teachers or test construction instructors to increase their skill of constructing test items and identify certain items which need revision. Heaton (1975) said that all items should be examined from the point of view of their level of difficulty (item difficulty) and discrimination (item discrimination).

Item difficulty (or the index of difficulty, or the facility value) is a measure of the difficulty of an item (Shakil, 2008). It is simply the percentage or proportion of the students taking the test who answered the item correctly (Matlock-Hetzel, 1997). The larger the percentage getting an item right, the easier the item. It can be calculated by using formula:

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

F.V.: the facility value (index of difficulty)

R: the number of correct answers

N: the number of students taking the test

Shakil (2008, p. 8) classified the ideal level of difficulty for multiple choice test items as follows:

Table 1. Classification of the ideal difficulty level for multiple choice test items

Number of alternatives	Ideal item difficulty level
2	0.75
3	0.67
4	0.63
5	0.60

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 (Assessment, 2005). A good test item should discriminate between those who have high score on the test and those who have low score (Backhoff, Larrazolo, & Rosas, 2000). The higher the discrimination index, the better the item can determine the difference between those with high test scores and those with low ones. It can be calculated by using formula:

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

D: Discrimination Index
n: number of candidates in one group
U: Upper group
L: Lower group

Ebel & Frisbie (1986) classified the index of discrimination as follows:

Table 2 Discrimination power according to the D value

D=	Quality	Recommendations
> 0.39	Excellent	Retain
0.30-0.39	Good	Possible to improve
0.20-0.29	Mediocre	Need to check / review
0.00-0.20	Poor	Discard/review in depth
< -0.01	Worst	Definitely discard

However, some best practices in item and test analysis are too infrequently used in actual practice (Matlock-Hetzel, 1997). Nowadays teachers rarely conduct an item analysis which involves item difficulty and item discrimination of their test items after giving test to the students (Azwar, 2000). It also happens in English teachers' of SMA Negeri 3 Jombang where the researcher conducted this study. Because of that, the researcher wants to conduct a research in analyzing the test items whether the test items that the teachers made are good or not based on the validity, reliability, the level of difficulty and also the level of discrimination.

For that reason, the present research is proposed to study the following problem: 1) How is the content validity of English final test for the first semester of the eleventh grade students of SMA Negeri 3 Jombang? 2) How is the reliability of English final test for the first semester of the eleventh grade students of SMA Negeri 3 Jombang? 3) How is the index of difficulty of English final test for the first semester of the eleventh grade students of SMA Negeri 3 Jombang? 4) How is the index of discrimination of English final test for the first semester of the eleventh grade students of SMA Negeri 3 Jombang?

METHOD

Referring to the research questions mentioned before, this study deals with the analysis of test items in the validity, reliability, item difficulty and item discrimination. This study is descriptive research because the researcher will only describe the quality of the test based on the result of the analysis in the form of written words without giving any treatment. Besides, the researcher also used quantitative approach in this study because it uses the numerical method to calculate and analyze the students' score interpreted with the items itself. Because of requiring descriptive and quantitative approach, the researcher used descriptive statistics to describe the data as the result of the study.

The setting of the study is SMA Negeri 3 Jombang. It is located at Jl. Dr. Sutomo No. 56 Jombang. For this study, the researcher chose the eleventh grade students as the subject of the study. Specifically, the sample of the study is XI IPA 5 class. The researcher chose the sample randomly.

There are five data of the study. All of the data are documents. The first data is the test items of the English final test as the main source of the data which will be analyzed. The second data is the Standar Isi of English in curriculum 2006 for the eleventh grade students for the first semester within the indicators from the teacher. Those two data will be used to analyze the validity of the test by matching the test items with the Standar Isi and the indicators itself. The third data is the students' score in which it will be used for analyzing the reliability. The steps are 1) making a tabulation of the students' scores orderly from the highest into the lowest, 2) measuring the mean of the students' scores, 3) measuring the standard deviation, and 4) measuring the coefficient of reliability. The fourth data is the students' answer sheets. The last data is the answer's key of the test items. Those two data will be collaborated with the main data, the English final test items, to calculate the index of difficulty and the index of discrimination of the test. The steps are 1) arranging the students' scores from the highest to the lowest score, 2) finding out the upper and lower group by dividing the class into two groups in equal size (the top half and the bottom half), 3) finding out the number of the students in upper and lower group who answer the items correctly in each item, 4) calculate the index of difficulty and discrimination. The researcher only took the data from the teacher because all of the data in the form of documents.

RESULT AND DISCUSSION

Result

The result of content validity showed that there are fifteen items in listening skill but there are only two items which is appropriate to the objectives. They are items number 3 and 5. The percentage of those items is 4%. Instead of both of them, the items are not appropriate to the stated objectives. There is an item which is fit to the objectives in speaking skill. It is item number 50. The percentage is 2% only. There are twenty-eight items which is appropriate to the objectives of reading skill (56%); eight items for responding report text, items number 16, 17, 18, 19, 20, 21, 22, and 23 (16%); ten items for responding narrative text, items number 29, 30, 31, 32, 33, 34, 35, 36, 37, and 38 (20%); ten items for responding analytical exposition text, items number 24, 25, 26, 27, 28, 39, 40, 41, 42, and 43 (20%). There are three items which is fit to the grammar. They are items number 45, 46, 47, and 49 (8%). There is no item for writing skill (0%). There are two items which are not appropriate at all such as items number 44 and 48 (4%). From the result, it can be concluded that 70% of the test items are developed based on the English curriculum of 2006 for the first semester of eleventh grade students even though there are 30% left which are not appropriate to that curriculum.

The result of reliability showed that the coefficient of reliability of the English test for the first semester of eleventh grade students of SMAN 3 Jombang 2012/2013 is 0.633 in which it belongs to adequate reliability because the teacher-made tests are usually considered as a test within adequate reliability if the reliability of the test is 0.60 or above. The mean score of the test is 37. There are 20 students who get score 37 or more than the mean score or the percentage is about 59%. The standard deviation (s.d) of the students' score is 5.04.

The result of item difficulty can be seen in the table as follows:

Table 3. The result of analyzing the item difficulty

Criteria of facility value	Number of item	Total
Very difficult (0.00 – 0.20)	41	1
Difficult (0.21 – 0.30)	26, 44	2
Moderate (0.31 – 0.70)	3, 9, 10, 30, 35, 38, 40, 45, 46, 47, 49	11
Easy (0.71 – 0.90)	5, 6, 7, 8, 11, 13, 17, 18, 19, 20, 21, 22, 23, 24, 29, 31, 32, 33, 34, 36, 37, 39, 48, 50	24
Very easy (0.91 – 1.00)	1, 2, 4, 12, 14, 15, 16, 25, 27, 28, 42, 43	12

Based on the table above, it showed that 11 items belong to moderate items which have appropriate or good level of difficulty such as item number 3, 9, 10, 30, 35, 38, 40, 45, 46, 47, and 49. Among those items, there are 5 items which are closely considered as having ideal level of difficulty (0.60). They are item number 3, 9, 30, 46, and 47. There are 24 items considered as easy items because they have difficulty level between 0.70 and 0.90. More than easy, there are 12 items belong to very easy items in which they have level of difficulty above 0.90. There are only 2 items which belong to difficult items such as item number 26 and 44 where they have the range level of difficulty between 0.20 and 0.30. The rest of those items, which is number 41, belong to very difficult item because it has the level of difficulty less than 0.20.

The result of the index of discrimination can be seen in the table as follows:

Table 4 The result of analyzing the item discrimination

Criteria of D value	Number of item	Total
Excellent (> 0.39)	3, 37, 40, 46, 49	5
Good (0.30 – 0.39)	18, 24, 38, 47	4
Mediocre (0.20 – 0.29)	5, 6, 7, 8, 11, 13, 19, 20, 21, 23, 32, 33, 36, 45, 50	15
Poor (0.00 – 0.20)	1, 2, 10, 12, 14, 15, 16, 17, 22, 25, 27, 28, 29, 31, 34, 35, 39, 42, 44, 48	20
Worst (< -0.01)	4, 9, 26, 30, 41, 43	6

From the table above, it showed that most of the items in English final test for the first semester of eleventh grade students of SMAN 3 Jombang 2012/2013 are poor. The poor items are 20 items or it reached 40% of the whole test items. Worst than the poor items, there are 6 items belong to worst items. However, there are 15 items (30%) which are on mediocre quality test items but it still needs to check or review. Besides, there are 4 items belong to good quality such as item number 18, 24, 38, and 47. Unfortunately, there are only 5 items which are categorized as excellent items.

Discussion

The test is adequately developed based on the 2006 English curriculum because the test presents 70% of the content objectives of the curriculum. However, the administration for each skill is not evenly spread. The most items are requiring the reading skill or mostly a half of the whole test items (56%). The listening skill covers 4% only. Moreover, the speaking skill is only on 2% covering the objectives. The grammar items cover 8% of the test. Unfortunately, there is one language skill left which is not measured by the test, writing skill because the test provides objectives test in the form of multiple-choice items only. The agreement of the test is 70%. It is not more than 75% (high content validity) and also not less than 50% (low content validity) so it can be concluded that the English test for the first semester of eleventh grade students of SMAN 3 Jombang 2012/2013 has adequate or moderate content validity.

From the result of the analysis of reliability, the coefficient of reliability is 0.633. It can be concluded that the English final test for the first semester of eleventh grade students of SMAN 3 Jombang is considered adequate reliability because the teacher-made tests are usually considered as a test within adequate reliability if the reliability of the test is 0.60 or above.

For the index of difficulty, mostly a half of the whole items are easy items, precisely 24 items or 48% of the test items which have the index of difficulty between 0.70 and 0.90. Besides, there are 12 items (24%) which are very easy so the numbers of those easy and very easy items are 72%. It can be concluded that the English final test items for the first semester of eleventh grade students of SMAN 3 Jombang are mostly easy so that it needs to improve.

For the index of discrimination, the English final test for the first semester of eleventh grade students of SMAN 3 Jombang has mostly poor index of discrimination. There are 20 items which belong to poor items. Those items should be discarded or review in depth. The worst items are 6 items which should be definitely discard. Besides, there are 15 items which are mediocre and those still need to check or review. However, there are 5 excellent items which can be retained and 4 good items which have possibilities for improvement.

CONCLUSION AND SUGGESTION

Conclusion

Based on the result of the analysis before, the researcher pointed out this study to the following conclusions. The English final test for the first semester of the eleventh grade students of SMAN 3 Jombang has moderate content validity because the test represents three basic competencies such as listening, speaking, and reading. There is one basic competency left that is writing. Moreover, there are two items in the test which are not appropriate to the indicators or objectives in the English curriculum for the first semester of eleventh grade students.

In the analysis of reliability, the coefficient is 0.633. It can be concluded that the English final test for the first semester of the eleventh grade students of SMAN 3 Jombang has moderate reliability because the teacher-made tests are usually considered as a test within adequate or moderate reliability if the coefficient reliability of the test is 0.60 or above.

For the index of difficulty, the test shows 48% of the test items are considered as easy items because they have difficulty index between 0.70 and 0.90. Moreover, 24% of the test items are very easy items in which the index of difficulty is above 0.90. However, there are still 5 items which are closely considered as having ideal index of difficulty (0.60). The rest of the items are difficult items (4%) and very difficult items (2%). It can be concluded that most of the test item of English test for the first semester of the eleventh grade students of SMAN 3 Jombang are easy items.

In the analysis of item discrimination, it can be concluded that most of the items in English final test for the first semester of eleventh grade students of SMAN 3 Jombang 2012/2013 are poor items. The poor items are 20 items or it reached 40% of the whole test items. Worst than the poor items, there are 6 items belong to worst items. However, there are 15 items (30%) which are on mediocre quality test items but it still needs to check or review. Besides, there are 4 items belong to good. Unfortunately, there are only 5 items which are categorized as excellent items.

Finally, the English final test for the first semester of the eleventh grade students of SMAN 3 Jombang has moderate content validity and reliability. The test mostly has easy item difficulty, and poor index of discrimination.

Suggestion

From the result of the analysis of this study, the researcher wants to give some suggestions in order to make better improvement for the test makers then. For the test makers, it is very important to construct good quality test items. The test makers have to know their purpose of giving the test. They should really know what they are going to measure. They have to construct each item based on the objectives in the curriculum stated in the Standar Isi. They should understand the kind of objectives should be achieved in the test. To make the test appropriate to the basic competencies, they should construct the test items based on each basic competency's objective stated in the curriculum. The test makers can learn how to construct good test items from the reference books especially the books which consist of principles in constructing a test within good index of difficulty, and good level index of discrimination in which each item will be able to differentiate students who get high and low scores. The test makers should construct the test based on the principles of constructing good test items so that the test will be acceptable and the objectives will be achieved. The test makers should try out the test items first before they give the test to the students to make sure whether the test items they made are already good or not. The test makers should revise or maybe remove the test items which are not suitable to the objectives in the curriculum, which are not appropriate to the good level index of difficulty, and which has poor or worst index of discrimination.

REFERENCES

Assessment, O. o. (2005). *ScorePak: ITEM ANALYSIS.* Seattle: University of Washington.

Azwar, S. (2000). *Reliabilitas dan Validitas.* Yogyakarta: Pustaka Belajar.

Bachman, L. f. (1990). *Fundamental Considerations in Language Testing.* USA: Oxford University Press.

Backhoff, E., Larrazolo, N., & Rosas, M. (2000). The Level of Difficulty and Discrimination Power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa* , 1-16.

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

Heaton, J. B. (1975). *Writing English Language Tests.* Singapore: Longman.

Matlock-Hetzel, S. (1997). *Basic Concepts in Item and Test Analysis.* Austin: Texas A&M University.

Shakil, M. (2008). *Assessing Student Performance Using Test Item Analysis and its Relevance to the State Exit Final Exams of MAT0024 Classes.* Hialeah: Miami Dade College.

Weir, C. J. (1993). *Understanding and Developing Language Tests.* London: Prentice Hall International (UK) Ltd.