AN ANALYSIS ON THE ENGLISH MULTIPLE-CHOICE ITEM TEST FOR PRIMARY STUDENTS

Mina, Regina, Bambang Wijaya
English Program Education Study Program, Faculty of Teacher Training and Education, Tanjungpura University, Pontianak
Email:mhien_na@yahoo.com

Abstract: This research concentrates on specific questions regarding the item validity, the test scores reliability, and item analysis in order to provide information that will lead to the improvement of test items construction. A descriptive method is applied to describe and examine the data. The research findings show that the test has fulfilled the criteria of having content validity. However the reliability value of the test scores is 0.67 which is categorized as unreliable. Through the item analysis, there are 11 items that are in need of improvement which are categorized "too easy" for the difficulty and "poor" for the discriminability. It means that almost 40% of the test items need to be revised as the items do not meet the criteria.

Keywords: item test, validity, reliability, item analysis

Abstrak: Penelitian ini dilakukan untuk menjawab beberapa pertanyaan spesifik mengenai validitas dan reliabilitas sebuah tes soal serta analisis butir soal yang bertujuan untuk menyediakan informasi yang akan mengarahkan pada peningkatan dalam penulisan butir soal. Sebuah metode deskripsi digunakan untuk menggambarkan dan menguji data yang tersedia. Hasil penelitian menunjukkan bahwa test tersebut valid tetapi nilai skor tidak reliabel. Berdasarkan hasil analisis butir soal, terdapat 11 butir soal yang harus diperbaiki dan ditingkatkan kualitasnya karena butir-butir soal tersebut dikategorikan terlalu mudah untuk dijawab sehingga mempunyai tingkat daya pembeda yang rendah.

Kata kunci: **butir soal, validitas, reliabilitas, analisis butir soal**.

In teaching-learning process, assessment plays important role. It provides information to the teacher on the area to which learning outcomes have been achieved by a student and the area to which he or she has been effective as a facilitator of learning. Assessment identifies what students know, understand, can do and feel at different stages in the learning process.

There are several forms of assessments. They can be written form or oral. They are paper-and pencil format, performance formats, long-term activity formats

and personal communication formats. The usage of those assessments depends on what area or to what extent the teacher seeks about his/her students' outcomes.

The most common assessments used by teachers, especially in SD Tunas Bangsa, a school which the researcher chose for her research are paper-and-pencil formats and performance formats. In paper-and-pencil formats item or usually we said as written tests item, have two general categories: (1) objective items which require students to select the correct response from several alternatives or to supply a word or short phrase to answer a question or complete a statement; and (2) subjective or essay items which permit the student to organize and present an original answer. Objective items include multiple-choice items, true-false items, matching items and completion items, while subjective items include restricted response items and extended response items.

One of the summative assessments used by the most teachers to assess their students; knowledge and comprehension is multiple-choice test. Multiple-choice items are easy to score, but the problem is, this type of tests is difficult and time consuming to construct. It is common knowledge that the correct answers should be distributed evenly among the alternative positions of multiple-choice items, but there are many other important guidelines for constructing good items and generally must be well known and recognized by the teachers. The guidelines are fairly comprehensive list of recommendations for constructing multiple-choice test items, focus on content, structure and options of a multiple-choice test item.

To produce or construct a good test, specifically multiple-choice test, it needs to be considered about the criteria. It needs to be done an items analysis. There are two ways in analyzing test items, using qualitative analysis and quantitative analysis. Qualitative analysis is done before the test items administered to the students. This analysis is done in order to know whether the test items appropriate and based on the constructing items guidelines which focus on the content, construct and language (Mardapi, 2004). In quantitative analysis, it focuses on the reliability, the discriminating power and the difficulty levels and practicality of the test items which will be tested (Suryapranata, 2004). Quantitative analysis analyzes the result (students' scores) of the test. Quantitative analysis is done after the test items administrated to the students.

A good test must be valid and reliable. In constructing a good test, teacher is expected to be able to plan the test in the table of items specification. It can help teachers in arranging the specific objectives of the test along with its contents. The researcher was motivated to do this research not only by a firmly held belief that teacher, especially English teachers in SD Tunas Bangsa never conduct an analysis of English multiple-choice items test because of the limitation of time and the difficulty in conducting item analysis but also they tend to reuse some items in the test. Therefore, the findings will provide important information for the teachers and the researcher. And the important thing is that analyzing items test is part of continuing professional development for teachers.

Starting from that statement before and point of view, the researcher was interested on analysis of the English multiple-choice items test constructed as a summative assessment at the second semester for primary two students at SD Tunas Bangsa, Kubu Raya in academic year 2012-2013.

The researcher was interested in learning and knowing how a good test is, how are the validity, the reliability, as well as the discriminating power of the English multiple-choice items test as a summative assessment in second semester for primary two students of SD Tunas Bangsa, Kubu Raya in academic year 2012-2013.

One way to be sure that a test provides representative sample of the learning outcome is to use a table of specification. Gronlund (1977: 9) defines a table of specification as a twofold table in which the learning outcomes are listed on one side and the subject-matter topics on the other. In addition, (Gronlund (1977: 27) claimed that the purpose of the table is to provide assurance that the test will measure a representative sample of the learning outcomes and the subject-matter topics to be measured. The quality of the test will depend on how closely the test maker can match the specifications (Gronlund, 1977: 2).

A test is said to have good quality when it can meet the requirement of having validity and reliability. A test is said to be valid if it is extent to which the test measures what is supposed to measure (Henning, 1987: 89). Any given test then may be valid for some purposes, but not for others. Reliability, on the other hand, is the extent to which a measuring device is consistent in measuring whatever it measures. Similarly, Henning (1987: 73) said that reliability is present when an examinee's results are consistent on repeated measurement.

Gronlund (1977: 110) explained that one way of investigating the quality of a test is to examine test takers' responses to each of the test items in a process called item analysis. After a test has been administered and scored, it is usually desirable to evaluate the effectiveness of the items. This is done by studying the students' responses to each item. The item analysis procedure provides the following information:

1. The difficulty of the item.
2. The discriminating power of the item.
3. The effectiveness of each distractor.

Thus, item-analysis information can reveal if an item is too easy or too hard, how well it discriminates between high and low scorers on the test, and whether all of the distractors function as intended. Item analysis data provides further information for improving test items.

Before a test is used to measure individual ability, the test must be tried out to identify the items that are weak and need to be modified or removed. in other words, test publishers must test the test. The test must be tried out before test publishers use it to measure someone's abilities. Djiwandono (2008: 203) stated that the carrying out of various try out tests is aimed to gather various information about the weaknesses, unclear instructions or even mistakes that can be found. So the quality of the test items before the test administration can be upgraded through the administration of try out tests. Meanwhile, Arikunto (2006: 200) said that there must be several try out tests in order to obtain a standardized test of which test items quality is ensured. A try out test must be administered to the testees that share the typical characteristics with the testees that are targeted. Djiwandono (2008: 203) said that trying out is a test activity for sample testees that have the same characteristics with the real target of testees' characteristics.

The item analysis is designed to ensure that the items fairly cover the field or criteria, the level of difficulty of the items is appropriate and the test is reliable (Cohen, Manion, and Morrison, 2005: 324-325). A simpler explanation about item analysis is written by Arikunto (2006: 205) that item analysis is a systematical procedure that gives specific information about the test items that we develop. She also said that by doing item analysis, a researcher can get information about the weaknesses of a test item and 'clues' to do some improvement.

In sum, a test needs items that are neither too difficult nor too easy, and the items should discriminate between the higher and lower scoring students.

METHOD

This research applied descriptive method to the problems of the research. Marczyk, DeMatteo, and Festinger (2005: 209) informed us that descriptive statistics allow the researcher to describe the data and examine relationships between variables. The analysis of this research is documentation-based. Documentation is one of the ways in collecting data by analyzing the notes and documents that are available. It becomes important when we wish to investigate how a document is made or used. All the suitable documents are the 23 answer sheets of the English multiple-choice items test that was administered as a summative assessment at the second semester for primary two students of SD Tunas Bangsa, in the academic year 2012-2013 on Friday, 31 May, 2013, the English test paper, the table of specification and the school scope and sequence of English (learning outcomes) of SD Tunas Bangsa.

Population refers to all any well-defined of people, events or objects that the researcher wishes to investigate (Ary, Cheser, and Razaviah, 1979: 129). The population in this research is all the 30 multiple-choice items (four alternatives are supplied in each item) of the English test together with the items of etes specifications, answer sheets of the total number of primary two students and the table of school scope and sequence of English (learning outcomes) of SD Tunas Bangsa.

The researcher applied documentary evidence or sources technique for collecting data. Tool of collecting data refers to methodologies used to identify information sources and collect information during an evaluation. The researcher used documentary analysis as the tool of collecting data.

In an initial stage, as the preparation, the researcher collected all the suitable documents. Next stage is analysis data. In this stage, they are three main points covering:

Validity

The researcher used content validity to see how well the content of the instrument represents the entire universe of content which might be measured. It can be best examined by making a table which consists of school scope and sequence of English (learning outcomes) of SD Tunas Bangsa, the item of specifications and the item tests that are placed with the school scope and sequence of English (learning outcomes) to identify whether or not the English school scope and sequence of English (learning outcomes) covered by the test.

Reliability

The reliability of a measuring instrument is the degree of consistency with which it measures whatever it is measuring. In this research, the researcher measured the reliability of each items by using Kuder-Richardson 21 (KR 21). The formula of KR 21 is as follows:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{M_t(n - M_t)}{(n)(S_t^2)}\right)$$

where,

r_{11} = the KR 21 reliability estimate
n = the number of items in the test
M_t = the mean of scores on the test
S_t^2 = the variance of test scores

(Sudijono, 2008: 258)

To obtain M_t , the researcher summed the individual scores of the distribution and divided by the total number of scores in the distribution. This relationship is shown in the following formula:

$$M_t = \frac{\sum X_t}{N}$$

(Sudijono, 2008: 258)

To obtain S_t^2 , the researcher used the following formula:

$$S_t^2 = \frac{\sum x_t^2}{N}$$

(Sudijono, 2008: 254)

To obtain $\sum x_t^2$, the formula is as follows:

$$\sum x_t^2 = \sum X_t^2 - \frac{(\sum X_t)^2}{N}$$

(Sudijono, 2008: 257)

Sudijono (2008: 209) shared the interpretation of the KR 21 reliability estimate (r_{11}) is as follows:

1. If r_{11} equals to or higher than 0.70, the test is considered to be reliable.
2. If r_{11} lower than 0.70, the test is considered to be unreliable.

Gronlund (1977: 142) notes that a complete lack of reliability would be indicated by a coefficient of .00 and perfect positive reliability would be indicated by a coefficient of 1.00.

Item analysis

In this research, item analysis covered:

1. Level of Difficulty

Item difficulty is determined as the proportion of correct responses, signified by the letter "p". The formula for calculating item difficulty is:

$$p = \frac{B}{JS}$$

where,

p = index of item difficulty
B = number of students answering correctly

JS = number of students taking the test

(Arikunto, 2006)

Tuckman (in Henning, 1987: 50) explained that an item that is rejected is the one with a proportion of correct answers that is less than 0.33 or that exceeds of 0.67.

Table 1
Classification of Difficulty Indices

Difficulty Index	Classification
Less than 0.33	Too difficult
0.33 – 0.67	Moderate
More than 0.67	Too easy

2. Discriminating Power

The first step of computing item discriminability is to separate the highest scoring group and the lowest scoring group from the entire sample on the basis of total score on the test. The students with highest total scores are compared in their performance with the students with lowest total scores using the formula:

$$D = P_u - P_l$$

where,

D = the index of discrimination

P_u = the proportion in the higher group

P_l = the proportion in the lower group

(Crocker and Algine, 1986: 314)

The proportion of the higher group (P_u) can be obtained through the following formula:

$$P_u = \frac{B_A}{J_A}$$

where,

P_u = the proportion of the higher group

B_A = the number of correct responses in the higher group

J_A = the number of testees in the higher group

(Sudijono, 2008: 390)

Meanwhile, the proportion of the lower group (P_l) can be obtained through the following formula:

$$P_l = \frac{B_B}{J_B}$$

where,

P_l = the proportion of the lower group

B_B = the number of correct responses in the lower group

J_B = the number of testees in the lower group

(Sudijono, 2008: 390)

Henning (1987: 51) stated that the decision to employ the number of students in each two groups is based on the optimal size of each group that is 28 percent of the total sample. The range of discriminability is from zero to one. The higher it is, the better.

According to Sudijono, the following is the classification and interpretation of discriminability index:

Table 2
Classifications and Interpretations of Discriminability Indices

Discriminability Index	Classification	Interpretation
Less than 0.20	Poor	The item has low discriminating power
0.20 – 0.40	Satisfactory	The item has sufficient discriminating power
0.41 – 0.70	Good	The item has good discriminating power
0.71 – 1.00	Excellent	The item has high discriminating power

(Sudijono, 2008: 389)

The discriminating power of an item is reported as a decimal fraction; maximum positive discriminating power is indicated by an index of 1.00. This is obtained only when all students in the upper group answer correctly and no one in the lower group does. Zero discriminating power (.00) is obtained when an equal number of students in each group answer the item correctly.

3. The effectiveness of Each Distractor

If a distractor elicits very few or no responses, then it may not be functioning as a distractor and should be replaced with a more attractive option. A good distractor in any items of a multiple-choice test format is the one that can attract the examinees to pick it as a correct answer. (Sudijono, 2008: 410).

To know how well a distractor work is by computation of 5% of the total examinees number (Arikunto, 2008: 411)

FINDINGS AND DISCUSSION

Findings

The results of the test analysis are presented in order to answer the research questions. Those research questions are about the validity of the English multiple-choice test, the test reliability, and the item analysis covering the level of difficulty, the discriminating power and the effectiveness of each distractor.

The purpose of the content analysis is to examine how all test contents cover the materials listed in the table of items specification. In addition to the purpose, content analysis examines whether the content of the test fulfills the expectation that stated in school scope and sequence of English.

Based on the result of analyzing content validity, every aspect of the learning content is as follows:

1. There are 6 items for reading which focus on answering questions based on the text.
2. There are 10 items about the using of tenses. The tenses are simple present tense, simple past tense and present continuous tense.

3. There are 5 items for identifying questions words used in the short dialogues.
4. There are 3 items for identifying the correct adjectives used as comparison in the sentences.
5. There are 6 items for identifying the nouns used in the sentences and showed from the pictures given.

Next, in order to estimate the reliability of the English test scores of the English test in the second semester of SD Tunas Bangsa in the academic year 2012-2013, Kuder-Richardson (KR 21) reliability coefficient was calculated.

The formula of KR 21 is as follows:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{M_t(n - M_t)}{(n)(S_t^2)}\right)$$

where,

r_{11} = the KR 21 reliability estimate
n = the number of items in the test
M_t = the mean of scores on the test
S_t^2 = the variance of test scores

(Sudijono, 2008: 258)

To obtain M_t, the researcher summed the individual scores of the distribution and divided by the total number of scores in the distribution. This relationship is shown in the following computation:

$$M_t = \frac{\sum X_t}{N}$$
$$= \frac{524}{23}$$
$$= 22.83$$

To compute S_t^2, the researcher used the following formula:

$$S_t^2 = \frac{\sum x_t^2}{N}$$
$$= \frac{357.91}{23}$$
$$= 15.56$$

To obtain $\sum x_t^2$, the formula is as follows:

$$\sum x_t^2 = \sum X_t^2 - \frac{(\sum X_t)^2}{N}$$
$$= 12296 - \frac{524^2}{23}$$
$$= 12296 - \frac{274576}{23}$$
$$= 12296 - 11938.09$$
$$= 357.91$$

The value of "$\sum X_t$" is obtained by summing the numbers of items correctly answered by all the testees while the "$\sum x_t^2$" is the value of "$\sum X_t$" squared. (See Appendix 4)

Finally, the computation of the test scores reliability using the Kuder-Richardson (KR 21) formula is presented as follows:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{M_t(n - M_t)}{(n)(S_t^2)}\right)$$
$$= \left(\frac{30}{30-1}\right)\left(1 - \frac{22.83\ (30 - 22.83)}{(30)(15.56)}\right)$$
$$= \left(\frac{30}{29}\right)\left(1 - \frac{22.83\ (7.17)}{466.8}\right)$$
$$= (1.03448275862069)\left(1 - \frac{163.6911}{466.8}\right)$$
$$= (1.03448275862069)(1 - 0.3506664524421594)$$
$$= (1.03448275862069)(0.6493335475578406)$$
$$= 0.6717243595425939\ \ (r_{11} < 0.70 = unreliable)$$

The item difficulty index is calculated for each test item in the English test. The researcher listed the students' answers in a table. The "0" in the cells of the table indicates the testees who miss that item while the "1" indicated the testees who answer the items correctly. Then the researcher summed the total number of correct items of each testee. The calculation of the difficulty indices of the 30 test items of which value is obtained by having the number of the correct items divided by the total items number. Difficulty indices that are below 0.33 are classified "too difficult", those lying from 0.33 to 0.67 are labeled "moderate", and the classification "too easy" is addressed to those above 0.67.

The results of the calculation are: there are 70% of the items or 21 items categorized as "too easy", 30% of the items or 9 items categorized as "moderate" and none of the items categorized as "too difficult".

The item discrimination index can be used to help determine if the questions are missed by those who know the material or those who do not. The researcher took 28% of the testees for each group, the "higher" group and the "lower" group.

The researcher obtained the discriminating indices by subtracting the proportion of the lower group from the proportion of the higher group. The results of the calculation based on the item discrimination indices are followed: there are 40% of the items or 12 items have low discriminating power, 16.67% of the items or 5 items have sufficient discriminating power, 40 % of the items or 12 items have good discriminating power and 3.33% of the items or only 1 item has high discriminating power.

Last, the calculation of the effectiveness of each distractor is classified into its category based on a theory saying that a distractor has functioned well if it is chosen by the examinees at least 5% of the total number of examinees. The results are: there are 6 items which all the distractors didn't function well. No students chose the distractors. It means the items are too easy, all students chose the correct answers.

Discussion

Evidence has been collected for the validation study of the English test for summative assessment of primary two in SD Tunas Bangsa in the academic year 2012-2013. The purpose of this section is to answer the research questions on the

basis of that evidence. Moreover, the research questions are answered using the data from the analysis to provide a better understanding of the results.

Content validity

It is impossible for a test to cover all the skills and materials in English that are supposed to be measured. There are several forms of assessment to be chosen from and it depends on the skill we are testing. The analysis performed in this test shows positive results. The content of the English test covers all the lesson materials listed in the table of items specifications developed prior to the test items writing.

Reliability

The reliability for the English test scores was estimated with the Kuder-Richardson (KR 21) reliability coefficient. The reliability value of the test scores is 0.67 which is categorized as "unreliable", as the value falls below the accepted minimum value "0.70".

Usually, the reliability level of test scores is influenced by the number of items on a test and the spread of proficiency levels. The more items a test has the more reliable the test scores are. If the range of proficiency levels of students is small and skewed towards the higher scores, the reliability values are higher too.

Level of difficulty

The analysis on the difficulty level of the 30 multiple-choice test items of the English multiple-choice item test constructed as a summative assessment at the second semester for primary two students of SD Tunas Bangsa, Kubu Raya in the academic year 2012-2013 shows that there are 21 items that are classified *too easy*, and 9 items that are classified *moderate*.

Although the items with the item difficulty index from 0.70 to 1.00, which are classified "too easy" represent a majority and or all of the students can answer the question correctly, there may be other validity concerns. Are these questions answered correctly because of the quality of instruction and the students' preparation level? Or, are these questions easily guessed and not reflective of the stated outcome. This is where combining the item difficulty index along with the item discrimination index can be useful.

Having known the difficulty level, the following actions as further follow-ups might be taken. First, the test items that can be stored in an item bank are items that are classified as *moderate*. They can be reused as good quality items in the future test. Second, there are possibilities of follow-ups for the items that are classified *too easy*, eliminate them or revise them.

Discriminating power

The item discrimination index can be used to see if a question is answered correctly more by the students in the high scoring group and is missed more frequently by those students in the low scoring group. This accomplished by dividing the students into two groups, namely high scoring group and low scoring group.

The result of item discrimination index can range from -1 to 1. The interpretation of this index is that if everyone answered the question correctly the

score would be 0. If everyone in the high scoring group answered correctly and everyone in the low scoring group missed the question, the item discrimination index would be 1. Conversely, if everyone in the low scoring group answered the item correctly and everyone in the high scoring group missed the item, then item discrimination would be -1. When the discrimination index falls below zero, this means that the testees in the low scoring group do better on that question than those in the high scoring group.

The discrimination index should not be used as the only one indicator for a good test. As the example, when one question is missed by every student in the class. The item discrimination index for this question would be 0. If everyone in the class correctly answers a question, the item discrimination index will also be 0. By looking at the item discrimination index along with the item difficulty index, a picture starts to come into view of the validity of the questions.

The discrimination value for 12 items of the total 30 items is below 0.20 and should either be rejected or revised. The discrimination ability of 5 items is satisfactory with a value between 0.20 and 0.40 and are in need of some improvement. 12 items discriminate reasonably well with a value between 0.41 and 0.70, but could possibly be improved. There is 1 item remaining, which has the discrimination ability categorized "excellent" with the value ranging from 0.71 to 1.00.

Overall, the items are not good enough indicate the ability of testees with a further consideration that there is only 1 item that discriminates very well between stronger and weaker students.

Briefly, the results of the item difficulty and item discrimination analyses show that there are many too easy items in general, which seem to lower the discrimination ability of the items. Most of the moderately difficult items discriminate well and only 1 item moderately difficult with poor discrimination value and 1 item moderately difficult with excellent discrimination value.

Effectiveness of each distractor

A test developer should make sure that all the distractors are plausible meaning that the distractors are believable and appearing likely to be true. If one distractor is obviously ridiculous, that distractor is not helping to test and discriminate between students. An incorrect distractor that is more prominent than the correct distractors needs to be reviewed as the quality of the distractors influences testees' performances on a test item.

One way to study responses to distractors is with a frequency table that describes the proportion of students who select a given distractor. It is recommended to remove or replace distractors selected by a few or no students because students find them to be implausible. Moreover, distractors mainly influence the difficulty index and item discrimination values of multiple-choice items. Sudijono (2008: 411) suggests that a distractor can be said to have functioned well when it is chosen by the examinees at least 5% of the total number of examinees. It can be concluded that each distractor should have a percentage of at least 5%. If a distractor has a value below 5%, it should be revised. Almost all items of the English test contain at least one distractor with a value below 5%. Making

distractors more plausible (appearing likely to be true) might help to increase discrimination values of items.

CONCLUSION AND SUGGESTION

Conclusion

From the findings of the research, we can conclude that the English test as follow: the test is good in terms of validity, but it is not good in term of reliability. Some items need to be revised because of its difficulty, the discriminability and the effectiveness of the distractors.

Suggestion

Based on the conclusion above, the following are general suggestions concerning the English test of the primary 2 students to improve the quality of the test: (1) Teachers have to spend more time to check each item's construction before administering the test to the students. The ambiguous or tricky items, the poor directions influenced the result of the reliability. (2) It needs to be considered to review and revise items with very low item difficulty and very low discrimination ability. (3) A further action needs to be taken to revise distractors that don't attract many responses and are thus not plausible at all. Revise distractors that are incorrect but attract more responses than the correct distractor. (4) Those items which categorized good in difficulty, discriminability and distractors can be put in item bank and reused.

BIBLIOGRAPHY

Arikunto, S. 2006. *Dasar-dasar Evaluasi Pendidikan, edisi revisi*. Jakarta: PT. Bina Aksara.

Cohen, L., Manion, L. and Morrison, K. 2005. *Research Methods in Education*, 5th edition. London and New York: Routledge/Falmer Taylor & Francis E-Library.

Djiwandono, S. 2008. *Tes Bahasa: Pegangan bagi pengajar bahasa*. Jakarta: Indeks.

Gronlund, N. E. 1997. *Constructing Achievement Tests*, 2nd Edition. USA: Prentice – Hall.

Henning, Grant. 1987. *A Guide to Language Testing: Development, Evaluation, Research*. Los Angeles: Newbury House Publishers.

Marczyk, G., DeMatteo, D. and Festinger, D. 2005. *Essentials of Research Design and Methodology*. USA: John Wiley & Sons. Inc.

Mardapi, D. 2004. *Penyusunan Tes Hasil Belajar*. Yogyakarta: Program Pascasarjana UNY.

Sudijono, A. 2008. *Pengantar Evaluasi Pendidikan*. Jakarta: Raja Grafindo Persada.f