AN ANALYSIS OF THE ENGLISH SUMATIVE TEST ITEMS

RESEARCH ARTICLE

By:

RUSTAMAN ABDUL SALAM
F12110054

ENGLISH LANGUAGE EDUCATION STUDY PROGRAM
LANGUAGES AND ARTS EDUCATION DEPARTMENT
TEACHER TRAINING AND EDUCATION FACULTY
TANJUNGPURA UNIVERSITY
PONTIANAK
2015

AN ANALYSIS OF THE ENGLISH SUMMATIVE TEST ITEMS

RESEARCH ARTICLE

RUSTAMAN ABDUL SALAM
F12110054

Approved by:

Supervisor 1	Supervisor 2
Dr. Iwan Supardi, M. Appling	**Urai Salam, Ph.D**
NIP.196612261994031004	**NIP.197001111998031001**

Mengetahui,

Dekan of FKIP	Head of Language and Arts Department
Dr. Martono	**Drs. Nanang Heryana, M.Pd**
NIP. 19580513 198603 1 002	**NIP.197001111998031001**

AN ANALYSIS OF THE ENGLISH SUMMATIVE TEST ITEMS

Rustaman, Iwan, Urai Salam
English Study Program FKIP Untan
Email: rustamanabduls@gmail.com

Abstract: An appropriate summative test should consider the validity, which consist of content validity, level of difficulty and discriminating power, and reliability. The aim of this research is to find out those terms of quality for the eighth grade English Summative test in SMP Negeri 1 Sungai Raya.The research design in this research is descriptive quantitative which present and describe the data. The samples of this reaserch in the sumative test items. The research finding showed that the content validity was classified into invalid test due to the unavailability of the table specification. The mean score for the level of difficulty is 0.582 and categorized as moderate summative test, and the discriminating power mean score is 0.368 and categorized as a good summative test. The reliability mean score is 0.742 and categorized as a substantial summative test. In general the test was classified as a good summative test despite the unavailability of the table of specification.

Key words:EnglishSummative,Test Items

Abstrak: Sebuah test yang baik harus harus memenuhi tingkat validitas yang baik, yang terdiri dari validitas isi, tingkat kesukaran dan daya pembeda, serta reabilitas. Tujuan dari penelitian ini adalah untuk mencari tahu kualitas tersebut pada butir soal ujian akhir semester ganjil di SMP Negeri 1 Sungai Raya. Metode penelitian yang digunakan adalah deskriptif kuantitaf yang menampilkan data dan menjelaskannya. Sample dari penelitia ini adalah butir-butir soal pada ujjian akhir semester ganjil. Hasil penelitian menunjukan bahwa validitas isi test dikategorikan sebagai test yang tidak sah dikarenakan ketidak tersediannya kisi-kisi soal. Nilai rata-rata untuk tingkat kesukaran adalah 0.582 dan dikategorikan kedalam test yang sedang, sementara itu untuk nilai daya pembeda adalah 0.368 yang dikategorikan sebagai test yang bagus. Selanjutnya, test tersebut mempunyai tingkat reliabilitas yang kuku dengan nilai 0.742. Secara keseluruhan butir-butir soal ujian akhir semester tersebut dikategorikan sebagai test yang bagus meskipun tidak ada kisi-kisi soal.

Kata Kunci: ***Test Sumatif Bahasa Inggris,Butir Soal***

Evaluation has a very important role in teaching learning process. It is conducted in order to find out whether instructional objectives established by the teacher have been achieved or not by the students. Airasian and Russell (2008) state that "evaluation is the product of assessment that produces a decision about the value or worth of a performance or activity based on information that has been collected, synthesized and reflected on. It is a process of making judgment about what is good

and desirable". One of the important things to evaluate in teaching learning process that a teacher is demanded to construct test, which can measure the students' achievement. Heaton (1988),"test is constructed primarily as devices to reinforce learning and to motivate the students or primarily as a means of assessing the student's performance in the language. In the former case, the test is geared to the teaching that has taken place, whereas in the latter case the teaching is often geared largely to the test". Moreover, Hughes (2003) state that a test is intended to measure students' achievement and the degree of success of the teaching learning program. It will measure the students' knowledge and allow them to know their progress. On the other hand, it will help the teacher to adjust his/her instruction on daily basis.

Being aware of this importance of evaluation in teaching learning process, the teacher or the test maker needs to have information or even to do evaluation by constructing a test as a tool of evaluation. To assess whether the students have mastered the material given for one semester, the teachers, as the test makers, should do the evaluation by giving an achievement test, such kind of test is called summative test. Since the summative test covers a wide range of materials learnt, the teacher or test maker should construct a test well, because it is aimed to find out how well the students have achieved the instructional objectives of a course. Besides, the teacher is able to know which students have achieved the instructional outcomes and which ones have not. Therefore, since evaluation has an important role in teaching and learning activity, the teacher should do the analysis of the test given. This analysis could be done before or after the test given. According to the English teacher statements in SMP Negeri 1 Sungai Raya, they did not do the analysis before or after the test given. They also did not give the try out for the test. Besides that, teacher did not have the table of specification of the test, which is as a measurement of the content validity. On the other hand, by analyzing the test, the teacher will get useful information for the class discussion of the test; help the students to improve their learning and feed back to prepare a better test in future.

However, the test used in the evaluation of the teaching learning process might not be able to achieve its goals. It means that the test might be invalid, unreliable, too easy, or too difficult for the students. It can be happened because the teacher or the test maker does not consider the validity, which covers the content validity, level of difficulty and discriminating power of the test and reliability. Validity and reliability are two important characteristics of measurement and evaluation. Brown (2004) defined that there are five criteria for testing a test: practicality, reliability, validity, authenticity, and wash back. In this research, the writer only focuses on reliability and validity because both elements are more suitable related to the research problems above. Validity is an important characteristic of a test. If tests do not truly measure what it is supposed to measure, the result is not value."Validity is the extent to which inferences made from assessment results are appropriate, meaningful and useful in terms of the purpose of the assessment" (Groundland, in Brown, 2004: 185).

A good summative test should cover these two important aspect, validity and reliability, of a test. In this research, the writer anlayzed the validity which consist of content validity, level of difficulty and discriminating power, and reliability. Validity in language test depends on the linguistic contents of the test and the situation or technique used to test this content. The test should aim to provide a correct measure of the particular skill it intended to measure. A valid test will provide information about the students' achievement. The test used in teaching learning process is designed to measure students' achievement based on learning objectives. Therefore, there is a relationship among test validity and learning objectives.

Content validity is concerned with the teaching materials that have been learned by the student. Ross (2004) states that content validity refers to the extent to which a test measures a representative sample of subject-matter content and behavioral content from the syllabus, which is being measured. It refers to the correspondence between the test items and test indicators, which related to the instructional objectives. A good test should be constructed based on the teaching materials and instructional objectives, which represented in form of table specification. An appropriate procedure to evaluate the content validity of a summative test is to match between the test items and the instructional objectives.

Level of difficulty of an item simply shows how easy or difficult the particular item in the test. Level of difficulty is generally expressed as the fraction of the students who answered item correctly. It does not show the certain item is good or not but it just shows that the item is easy or difficult for the test taker or examinee. Level difficulty expresses the proportion of the students answering the test items correctly. The purpose is to make a difference between the test taker and students, to spread them out in term of their performance on the test. In another word, it is to find out whether the test items are categorized as revised, difficult, moderate, or easy.

Discriminating power of the test items is to measure how performance on one item correlates to performance in the test as a whole. It is the degree to which students with high overall exam scores also got a particular items correct. On the other words, discriminating power is to find out how well the test items separating the high group students from the low group students who answer the test items correctly. A good discriminating power is the upper group students answer the item correctly more frequently than the lower group students do. In some occasion, it is often found that the score of the discriminating power is negative which means there are more students from the lower group who answer correctly rather than the upper group. This items should be rejected and no need to use for the future test.

Reliability addresses the question of whether the results of measuring process are consistent on occasions when they should be consistent. Essentially, reliability sets an upper limit for validity. In other words, if a test is not reliable, there is a great deal of measurement error. If a test is highly reliable (little measurement error), it proves that the test has the potential to be highly valid or vice versa.

METHOD

The form of research design that was used in this research is descriptive quantitative research. It was used to describe what is, describing, recording, analyzing, and interpreting conditions that exist (Best & Kahn, 2006). By using a quantitative research, researcher gains a systematic calculation results about the contents of a document by using the numbers statistical results thus obtained the expected value or percentage. “Descriptive statistic describes and presents the data collected in the research study” (Cohen, Manion & Morrison, 2007: 503). In this research, the writer would give description and present the data.

The population of this research is English Multiple-choice Summative Test Items for first Semester of VIII Grade Students of SMP Negeri 1 Sungai Raya in Academic Year 2014/2015. The total numbers of the test items are 40, and there are 34 students’ answer sheets that are going to measure as a sample. In order to solve the problems objectively in this research, the writer used the document analysis to collect data. The researcher collected the data of related information including the result of VIII grade student test of the first semester. In gathering the necessary data,

the writer will collect the test administrated and scored by the teacher, the students' answer sheets, and the table of the specification that was given by the teacher. After that, the writer will analyze the data based on the problems designed: validity (content validity, level of difficulty and discriminating power) and, reliability.

To analyze the data of this research, the writer will take the data from the information about the summative test that mention above. In analyzing the validity, the writer will divide into two parts. First, to analyze the content validity, the writer will use the table of specification and will match the items and the indicators. Second, to measure the level of difficulty and discriminating power, the writer will use the Test Analysis Program (TAP) software. Furthermore, to analyze the reliability the writer will also use Master TAP. Finally the writer will insert the students' answer from the answer sheets to the program.

FINDINGS AND DISCUSSION

Findings

1. **Analysis of Content Validity**

In analyzing the content validity, the writer was going to use the table of specification. Unfortunately, there was no table of specification that provided by both the English teacher and the school. It made the test could not be valid in term of content because the writer could not measure the content validity without the table of specification. The unavailability of table of specification was caused by the absence of targeted grade representative of the school in the test composition process. The English teacher representatives of grade seven to grade nine from appointed schools in Kabupaten Kubu Raya constructed the test items. The table of specification was only given to schools whose representative teacher attended the test items composition. However, during test items composition, there were only two English teachers representing SMP Negeri 1 Sungai Raya, those are from grade seven and grade nine. There was no representative for the eighth grade of the school. Due to the absence, thus SMP Negeri 1 Sungai Raya did not get the table of specification for eight grade's summative test.

2. **Analysis of Level of Difficulty**

The writer used MasterTAP software to analyze the difficulty level of the test. The results of data analysis are as follow:

Table1
Item Level of Difficulty Analysis

Number of test items	Level of difficulty	Classification
1	0.50	Moderate
2	0.71	Moderate
3	0.79	Moderate
4	0.29	Revised
5	0.47	Difficult
6	0.74	Moderate
7	0.56	Moderate
8	0.00	Revised

9	0.91	Easy
10	0.68	Moderate
11	0.71	Moderate
12	0.82	Easy
13	0.71	Moderate
14	0.76	Moderate
15	0.88	Easy
16	0.12	Revised
17	0.76	Moderate
18	0.41	Difficult
19	0.97	Easy
20	0.76	Moderate
21	0.59	Moderate
22	0.76	Moderate
23	0.32	Difficult
24	0.76	Moderate
25	0.82	Easy
26	0.53	Moderate
27	0.50	Moderate
28	0.21	Revised
29	0.74	Moderate
30	0.68	Moderate
31	0.15	Revised
32	0.24	Revised
33	0.26	Revised
34	0.71	Moderate
35	0.44	Difficult
36	0.56	Moderate
37	0.50	Moderate
38	0.65	Moderate
39	0.68	Moderate
40	0.65	Moderate

a. There were 7 test items that classified as revised items. Those items were the item number 4, 8, 16, 28, 31, 32, and 33.
b. There were 4 test items that classified as difficult items. Those items were the item number 5, 18, 23, and 35.
c. There were 24 test items that classified as moderate items. Those items were the item number 1, 2, 3, 6, 7, 10, 11, 13, 14, 17, 20, 21, 22, 24, 26, 27, 29, 30, 34, 36, 37, 38, 39, and 40.
d. There were 5 test items that classified as easy items. Those items were the item number 9, 12, 15, 19, and 25.

Table 2
The Criteria of Level of Difficulty

Index of Level of Difficulty	The Qualification
Minus to 0.29	Revised (R)
0.30 to 0.49	Difficult (D)
0.50 to 0.79	Moderate (M)
0.80 to 1.00	Easy (E)

From those 40 items, the English summative test items for the first semester of eight grades in SMP Negeri 1 Sungai Raya in academic year 2014/2015 could be categorize as moderate. It was concluded by the mean of the level of difficulty with the score 0.582.

3. **Analysis of Discriminating Power**

To analyze the discriminating power the writer used MasterTAP software. The discriminating power shows the difference between students in the upper group and the lower group. The results of discriminating data analysis are as follow:

Table 3
Item Discriminating Power Analysis

Number of TestsItem	Discriminating Power	Classification
1	0.37	Good
2	0.68	Very Good
3	0.12	Discarded
4	-0.13	Discarded
5	0.89	Very Good
6	0.56	Very Good
7	0.58	Very Good
8	0.00	Discarded
9	-0.09	Discarded
10	0.67	Very Good
11	0.67	Very Good
12	0.44	Very Good
13	0.67	Very Good
14	0.33	Good
15	0.22	Sufficient
16	-0.01	Discarded
17	0.23	Sufficient
18	0.38	Good
19	0.11	Discarded
20	0.34	Good
21	0.47	Very Good

22	0.44	Very Good
23	0.38	Good
24	0.34	Good
25	0.23	Sufficient
26	0.28	Sufficient
27	0.68	Very Good
28	0.08	Discarded
29	0.44	Very Good
30	0.78	Very Good
31	0.09	Discarded
32	0.29	Sufficient
33	0.39	Good
34	0.34	Good
35	0.58	Very Good
36	0.36	Good
37	0.36	Good
38	0.34	Good
39	0.44	Very Good
40	0.37	Good

a. There were 8 test items that qualified as discarded items or poor items. Those were item number 3, 4, 8, 9, 16, 19, 28, and 31.
b. There were 5 test items that qualified as sufficient items. Those were item number 15, 17, 25, 26, and 32.
c. There were 12 test items that qualified as good items. Those were item number 1, 14, 18, 20, 23, 24, 33, 34, 36, 37, 38, and 40.
d. There were 15 test items that qualified as very good items. Those were item number 2, 5, 6, 7, 10, 11, 12, 13, 21, 22, 27, 29, 30, 35, and 39.

Table 4
The Criteria of Discriminating Power

Discriminating Power	**Qualification**
0.40 – 1.00	Very Good
0.30 – 0.39	Good
0.20 – 0.29	Sufficient
0.00 – 0.19	Discarded

Based on the test items analysis, it was found that there were 3 items of discriminating power with negative values. It means that there were more students from the lower group who answer the test correctly, rather than the upper group does. Those items were number 4 with the discriminating power value -0.13, item number 9 with the discriminating power value -0.09, and item number 16 with the discriminating power value -0.01. These items should be rejected. The mean score of the discriminating power was 0.368. It was categorized as good test.

4. **Analysis of Reliability**

The writer used MasterTAP software to analyze the reliability of the test items. Moreover, the writer took the Kuder-Richardson (KR 21) as a result. From the calculation it was found that the coefficient value of the test reliability was 0.742. Then, related to the table of reliability qualification, it could be concluded that the English summative test items in SMP Negeri 1 Sungai Raya in academic year 2014/2015 was qualified as substantial. The test was good, but could have been made without right procedure. In brief, regarding to the reliability result, the test had a good reliability, in general, even though there was no table of specification. It could be concluded that the test was made without right procedure.

Table 5
The Reliability Coefficient

Coefficient	Classification
0.00 - 0.20	Negligible (N)
0.21 – 0.40	Low (L)
0.41 – 0.60	Moderate (M)
0.61 - 0.80	Substantial (S)
0.81 – 1.00	High (H)

Discussion

A good test should be conducted based on the table of content or table of specification. The function of the table of specification is to measure whether or not the test covers all of the teaching materials. It is essential for a table of specification to present the lesson materials through indicators. The use of indicator is to measure the learning objective and students' achievement.

In this research, the writer supposed to do the content validity analysis but the school did not provide the table of specification. It was caused, the absence of table specification, due to some main factors. Based on the interview with the eighth grade English teachers in SMP Negeri 1 Sungai Raya, the first factor is that the test was conducted by the English teacher representative from some junior high schools in Kabupaten Kubu Raya. These English teachers were assigned by the education authority randomly. They were divided based on their level of teaching. In fact that the English teacher in SMP Negeri 1Sungai Raya eighth grade was not assigned to participate in conducting the summative test.

The second factor is that the school did not have the table of specification due to there is no representative during the test conducting. The table of specification is only given to the school that has a representative as the test maker. Based on the evidence, there is no clear clarification why, those school that has no representative as a test maker, did not get the table of specification. Furthermore, the English teacher itself did not know which schools' representative made the test. Based on those two evidences above, the writer concluded that the test could notbe said as a valid or appropriate test in terms of content validity.

Based on the analysis of level of difficulty, the result showed that the mean score for difficulty level was 0.582 with minimum item difficulty level was 0.000 and maximum level of difficulty was 0.971. The data showed that were 7 test items classified as revised item. These items should be revised whether the quality of the question or the option if the test is going to use for the future. Moreover, 4 test items were classified as difficult item need to be reduce the difficulty level. The 24 test items that classified as moderate items were appropriate enough for the test. Finally, for the5 test items that classified as easy item need more improvement to reach the level of difficulty standard, consequently those items could be used for the next summative test. In brief, the level of difficulty of the summative test was 0.582 which mean that the test was moderate, but need revision for some items.

The result for discriminating power showed that the mean score was 0.368 which categorized as good summative test item. In addition, the minimum score was -0.133 and the maximum score was 0.889. The test analysis found that there were 3 items of discriminating power with negative values. It means that, there were more students from the lower group who answer the test correctly, rather than the upper group students do. These items should be rejected and no need to use for the future summative test. Furthermore, the mean score did not clarify that all of the test items were categorized as good discriminating power items. There were still need improvements for some test items for a better test.

The reliability for the English summative test of the eighth grade for first semester at SMP Negeri 1 Sungai Raya in academic year 2014/2015 was analyzed by using MasterTAP software. The reliability for KR21 was 0.742 which was categorizes as substantial. It means that to obtain a KR21 Reliability of 0.90 or categorized as high to very high reliability, the test must be 2.18 times longer, for a total of 87 items of similar quality to those in the test. Furthermore, the test was consisted of 40 test items, all included, 34 examinees, 14.00 minimum score, 34.00 maximum score, 22.00 median score and 23.294 mean score.

CONCLUSION AND SUGESTION

Conclusion

Based on the results that mention above, it could be concluded that the English Summative test for the first semester of eight grades at SMP Negeri 1 Sungai Raya in academic year 2014/2015 was classified as a good summative test item in term of level of difficulty, discriminating power, and reliability. In contrast, the content validity was categorized as an invalid summative test. Finally, for a better summative test, both the English teacher and the Education authorities should do the right procedure of making a test through conducting and providing a table of specification for the summative. This table of specification should be conducted based on the indicators of the test items which related to the basic competence and standard competence in the syllabus.

Suggestion

From the conclusion above, the writer would like to give some suggestions to conduct a better English summative test in the future as follows: (1) It is suggested, in conducting a test, that the test maker or the teacher should also compose a table of specification of the test. It is not only to help the teacher but also help the school to

measure and determine which learning objective has been achieved by the students. (2) The teacher should spread or give the table of specification to the students to help them in preparing the test. By having the table of specification, the students will be able to figure out which topic or teaching material that will be tested in the summative test. (3) The Education authorities and the government in Kabupaten Kubu Raya should hold constructing-test training for all of the active English teachers.

REFERENCES

Airasian, P. W & Russel, M. K. (2008). *Classroom Assessment: Concept and Evaluation, Sixth Edition*. New York: McGraw-Hill Companies.

Best, J. W. & Kahn, J. V. (2006). *Research in Education, 10th Edition.* Boston: Pearson Education.

Brown, H. D. (2004). *Language Assessment Principles and Classroom Practice*. California: San Francisco State University.

Cohen, L., Manion, L. &Morrison, K. (2007). *Research Methods in Education*. New York: Routledge.

Heaton, J.B. (1988). *Writing English Language Test, A practical Guide for Teachers of English as a second or foreign language*. Longman.

Hughes. A. (2003). *Testing for Language Teachers, Second Edition*. Cambridge: Cambridge Press University.

Ross, K. N. (2004). *Educational Research: Some Basic Concepts and Terminology*. Paris: International Institute for Educational Planning/UNESCO.