ENGLISH LISTENING TEST ITEMS EVALUATION: A CASE OF A TEACHER-MADE TEST OF SMK N 5 PONTIANAK 2014

Windy Ayu Lestari, Ikhsanudin, Eusabinus Bunau
English Education Study Program, Languages and Arts Department, Teacher Training and Education Faculty of Tanjungpura University in Pontianak
Email: Windypost14@gmail.com

Abstract

The purpose of this research is to evaluate and provide information about the quality of English listening test items. Evaluation a set of test items are to measure the content validity, the reliability, the level of difficulty, the discriminating power, and the distracters, for the first semester of grade twelve students in SMK N 5 Pontianak in Academic year 2013/2014. This research is an evaluation to a set of teacher-made test that consist of 15 multiple-choice test items. There are 31 students' in one class as the participant of the test. The data of this research were collected using a documentary technique. The data were taken from the result of an English listening summative test, English listening test paper, answer key, students' answer sheets, listening script, and test item specification. The finding of this research is the content of all of the test items is valid based on the test item specifications. However, based on the criteria used to classify the degree of reliability, the items are not reliable; the score is only 0.183 (negligible). Although the content of all of the test items is valid, the test items cannot be used continuously because it is not reliable, and the test items need several revision. Furthermore, the mean of items difficulty level has fulfilled the requirements of good test items in term of difficulty level with 0.688 (Moderate). The mean of discriminating power is 0.256 (moderate) that means it has fulfilled the requirements of good test items. There were 26 distracters that should be revised. Besides, there are 4 good test items which still can be used as reference for the next English summative test. There are 3 test items that should be discarded or changed by the other test items and 8 test items should be revised if the teacher wants to use the test for the next English summative test.

Keywords: Item Analysis, Listening Test Items

INTRODUCTION

Test is a tool of measurement. Testing in education is one of the important ways to measure the students. It is an attempt to measure a person's knowledge, intelligence, or other characteristics in a systematic way. (Fulcher, 2010) stated that "the purpose of such testing is primarily related to the needs of the teachers and learners working within a particular context." The purpose of giving tests is to discover the learning abilities of the students, to plan future instruction toward the students and to see how well teachers' teaching learning strategy or method. Through testing, the teacher can measure students' learning process.

This research is focus on teacher-made test. Through this research, the teacher can figure out, how to measure the language teaching test such as listening, speaking, writing, and reading in term of validity, reliability, level of difficulty, discriminating power and distracters. But, the researcher tends to focus on summative test for listening section as final exam for grade twelve students because listening test

is more than hearing the words or sentences. It is memorizing, thinking and analyzing what the listeners have heard. According to (Vandergrift, 2004) listening has gained much attention both in research and in language pedagogy as it has changed its role from a passive activity which deserved less class time to an active process through which language acquisition takes place. And among four language skills in teaching English, listening skill is used most frequently. According to (Feyten, 1991: in Nichols and Leonard, 1957: Rankin, 1930) in daily communication, people spend 45% of time in listening, 30% on speaking, 16% on reading, and only 9% on writing. Then listening is occupies an important role in learning process.

Based on the pre-observation done in SMK N 5 Pontianak, it was found out that the English teacher constructed the test by herself for students' examination. The teacher constructed the test paper and the audio for listening test section. As the teacher said when the interview was administrated, the test was used directly after the construction without any try out, eventhough there was a listening practice in teaching learning process. The test was administrated on December 4th 2014, after the test administrated the teacher analyzed the test by using ana-test application just for teacher references about their student achievement not for the test analysis in specific. Hence, the researcher was interested in the evaluation of the content Based on the statement above, the researcher take the data in this research from:

The tools of data collecting are from observation and analysis. The researcher find out the information related to the test, research problems and purposes in SMK N 5 Pontianak through interviewing the teacher and observe the school. The researcher get the data of the English test paper, the students answer sheets, the answer key, the listening test script and the form of test specification. Then analyze the test and the data that being collected by the teacher based on content validity, reliability, level of difficulty, discriminating power, and distracters. The researcher used the documentary analysis techniques to collect the data. The researcher collected the data by using the document of the related information, such as English listening test paper, students' answer sheet, the answer key, listening script test, and test specification of the test items.

validity, the reliability, the level of difficulty, the discriminating power, and the distracters of English listening test items for grade twelve students at SMK N 5 Pontianak in academic year 2013/2014. The purpose of the evaluation is for test improvement and quality assurance and also to show the error of the test. This research is not to judge the teacher who made the test, but to help the teacher to find out the quality of the test for better future test and to reach better result for analyzing student achievement.

METHOD

This research has the purpose to evaluate the object of research based on the data. As (Silver, 2004) stated that "...evaluation is a process of acquiring information. Evaluation of an innovation or an activity, a curriculum or organisational change, raises a series of sometimes difficult or contentious issues". (Gall et all, 2007) emphasized that "Educational evaluation is the process of making judgments about the merit, value, or worth of educational programs." In this research the researcher concern is to evaluate the test items.

The population in this research is refers to the test items. The population is 15 multiple-choice of the test items with four alternatives. And, there are 31 students' of grade twelve in one class as the participants who answering the test.

a. 15 multiple-choice questions of the test items.

b. 31 students' answer sheets.

First of all, the researcher took the data from observation in SMKN 5 Pontianak. Next, the tests were administered and scored by the teacher. The researcher collected the data. Then the researcher analyzed the data based on the problem designed: content validity, reliability, level of difficulty, discriminating power, and the distracters of the test items. At the end, the researcher make a report as the result of the test evaluation.

Furthermore, the content validity is concerned with the materials that the students have learned. (Hughes, 2003) suggests that "In order to judge whether or not a test has content validity, we need a specification of the skills or structure, etcetera". A comparison of test specification and test content is the basis for judgments as to content validity. Therefore, to measure that the test has high content validity the researcher needs a table of test specification. The table indicates the materials that the teacher or the test maker wants to test and this table should have been constructed before administering the test. Hence, in this research the researcher analyzed whether the test items are suitable or not with the materials that have been learned by the students.

The criteria to classify the level of difficulty:

Difficulty Index	**Classification**
Less than .30	Too Difficult
.30 to .70	Moderate
More than .71	Too Easy

Source: Thorndike and Hagen (as cited in Fiktorius, 2014)

According to (Gronlund, 1977) to estimate item discriminating power by comparing the number of students in the upper and lower groups who answer the item correctly. The discriminating power of an item is reported as a decimal fraction; maximum positive discriminating power is indicated by an index of 1.00. The discriminating power (DP) calculates by using TAP (Test Analysis Program) application to find out the discriminating power of the test directly.

The criteria using to determine the discriminating power:

Index of D	**The Qualification**
0.00 – 0.19	Revised
0.20 – 0.29	Moderate
0.30 - 0.39	Good
0.40 - 1.00	Very Good/Exellent

Source: Best. (2006)

Then, the reliability is element that determines the quality of our measurement instruments. According to (Airasian, 2000) reliability refers to the stability or consistency of assessment information and reliability is not concerned with the appropriateness of the assessment information collected. In this research, to measure the reliability of the test items, the researcher uses Kuder-Richardson 20 (KR 20). The reliability calculates by using TAP (Test Analysis Program) application to find out the reliability of the test directly. The following criteria used to classify the degree of reliability are: (a)Coefficient/r 0.0 – 0.20 is negligible. (b)Coefficient/r 0.20 – 0.40 is low. (c) Coefficient/r 0.40 – 0.60 is moderate. (d) Coefficient/r 0.60 – 0.80 is substantial. And (e) Coefficient/r 0.80 – 1.00 is high to very high. (Source: Best & Khan, 2006).

Afterwards, the level of difficulty of item shows how easy or difficult the particular item proved in the test. The difficulty of a test item indicates the proportion of test takers who answered correctly. The level of difficulty (LD) calculates by using TAP (Test Analysis Program) application to find out the level of difficulty of the test directly.

And then, a good distracter will attract more students and distracters are termed not useful if they are not selected by any students at all. Miller et al (2009) emphasized that any distracters that are not chosen by any test takers are poor distracters. If the distracters are poor it should be eliminated or replaced with a more attractive or plausible option. And the other hand, a distracter can be claimed to Item No. 1

Distracters A with 26 examinee has 26 / 31 x 100% = 83,87 % effectiveness index as the answer key.

Distracters B with 3 examinee has 3 / 31x 100% = 9,67 % effectiveness index as the good distracter.

Distracters D do not function well, while distractor B and C does. Specifically, distracter D is not chosen by any examinee and it is simply not contributing to the quality of the item at all, so it should be eliminated or change.

FINDING AND DISCUSSION

Findings

In getting the result of content validity analysis, the content validity is assessed by an evaluative technique using the content validity form. The form or the test specification is matching with the test itself in case to find out whether the test and the form are appropriate or not. Based on the data shown, it is concluded that the test has fulfilled the criteria of having the content validity.

In getting the result of analysis of the reliability the researcher used Kuder-Richardson formula (KR-20) the data was calculated by applying Master TAP (Test Analysis Program). In this case the researcher using master TAP version 4.2.5 is by Gordon P. Brooks in 2002. From the calculation through TAP it is found the coefficient of test item reliability is 0.183. Based on the classified coefficient of the test item reliability, the test items are Negligible.

function well if it has a strong power of attracting that it is chosen by at least 5.00% of the test takers, (Anderson & Morgan, 2008). The computation of how well a distracters works by computing how many students answer each choice (e.g A / B / C / D) that divided with the number of the examinee X 100%. The result will show us the percentage of each choice. For instance:

Distracters C with 2 examinee has 2 / 31 x 100% = 6,45% effectiveness index as the good distracter.

Distracters D with 0 examinee has 0 / 31 x 100% = 0 % effectiveness index as the poor distracter.

Data analysis of item difficulty level was computed by using the Master TAP application. The result showed mean of item difficulty is 0.688 (moderate). Furthermore, the result showed 9 too easy test items, 1 too difficult test items and 5 moderate test items. Based on the data calculation, the difficulty level of the test items which need revision and categorized as difficult, moderate and too easy items as follows:

(a)The item which categorized as too difficult is the item number 3. (b)The items that belong to the moderate test items are the items number 4, 8, 9, 11, and 12. (c)The items which categorized as too easy are the items number 1, 2, 5, 6, 7, 10, 13, 14, and 15.

Data analysis of the discriminating power was computed by using the Master TAP application. The result showed mean item discrimination is 0.256 (moderate). The calculation found 4 excellent test items, 3 good test items, 4 moderate test items and 4 revised test items. From the calculation of discriminating power the items which are belong to revised, moderate, good and very good as follows:

(a)The items that belong to the excellent items are items number 4, 8, 12 and 13. (b)The items which categorized as a good test items are items number 2, 11,

and 15. (c)The items that classified as moderate test items in discriminating higher and lower student are the items number 1, 3, 6, and 14. (d)The items number 5, 7, 9 and 10 were classified as poor or revised test items.

Each percentage of the calculation or the distracter effectiveness index is classified into its categories, that is poor and good based on a theory arguing that a distracter is claimed to function well is the one chosen by at least 5.00% of the total number of examinees, Anderson & Morgan (2008). Hence, the 15 items comprise 19 good distracters, 26 poor distracters, and 15 answer keys.

Discussion

A good test items should be based on the table of test specification in term of the content in purpose the materials are not too much exist in the test besides the other materials are less.

The results showed: (a)Questions number 1,2 and 3 are using a pictures. Based on the indicator, question number 1, 2 and 3 are suitable with the indicators. (b)Questions number 4, 5 and 6 are using questions-responses. Based on the indicator, question number 4, 5 and 6 are suitable with the indicators. (c) Questions number 7, 8, 9 and 10 are using Short Conversations. Based on the indicator, question number 7, 8, 9 and 10 are suitable with the indicators. (d)Questions number 11, 12, 13, 14 and 15 are using Short Talks. Based on the indicator, question number 11, 12, 13, 14 and 15 are suitable with the indicators.

The result of the evaluation is the content validity of the test items is valid. All of the items is match and based on the test specification. Hence, there is no problem with the content validity of the test.

From the calculation by using TAP, it is found out:

(a)The minimum score is 6,000 with 40,0 %. (b)The maximum score is 14,000 with 93,3%. (c)The median score is 10,000 with 66,7%. (d)The mean score is 10,323 with 68,8%. (e)The standard deviation is 1,654. (f)The skewness is -0,136. The skewness is minus, it meant the test that showed through bar chart or histrogram is inclining to the right. (g)The kurtosis is 0,103. It is showed the curva slope. (h)The potential problem items is 12 items such as items number 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14 and 15

The coefficient of test items reliability is 0.183. Based on the classified coefficient of the test item reliability, the test items are Negligible.

The result of the test items reliability is negligible and a good reliability is should be 0.60-above. The test can't be used continuously because the test item is not reliable. Hence, the test items need several revisions.

The research finding shows that some of the items were not fulfill the requirement of good test because they are too easy or too difficult. Item number 1, 2, 5, 6, 7, 10, 13, 14, and 15 are too easy. Item number 3 is too difficult, there were only 4 students who can answer the item. The statistical result showed the level of difficulty is 0.688 (moderate).

A good level of the test is in 0.30-0.71 (moderate). Based on the criteria and the result, some items need revision. And the statistical result showed the Discriminating power is 0.256 (moderate).

The items number 5, 7, 9 and 10 were classified as poor or revised test items so it is unable to discriminate upper and lower group students. The items fail to discriminate the upper group and the lower group. They are affected by the ineffective distracters are not plausible and attractive to the uninformed that enable students select the correct answer and eliminate those incorrect alternatives.

A distracter can be claimed to function well if it has a strong power of attracting that it is chosen by at least 5.00% of the test takers, Anderson & Morgan (2008). There are 26 distracters should be revised because

was not choose by at least 5.00% of the test

(a)Revised Distracters for item number 1 are C and D. (b) Revised Distracters for item number 2 are A and D. (c) Revised Distracters for item number 3 is B. (d) Revised Distracters for item number 4 are C and D. (e) Revised Distracters for item number 5 are all distracters. (f) Revised Distracters for item number 6 are A and D. (g) Revised Distracters for item number 7 are all distracters. (h) Revised Distracters for item number 8 are C and D. (i) Revised Distracters for item number 9 are A and C. (j) Revised Distracters for item number 10 are A and D. (k) Revised Distracters for item number 11 is no need. (l) Revised Distracters for item number 12 is A. (m) Revised Distracters for item number 13 is D. (n) Revised Distracters for item number 14 are B and D. (0) Revised Distracters for item number 15 is C.

Furthermore, as the result of the analysis the writer divided the test items into three groups. They are revised test group, bad test group and good test group. The revised test is a test which has revision in one of the analysis factor whether the level of difficulty or the discriminating power of the test. There are 8 test items which is included to this group. The test items numbers are 1, 2, 3, 6, 9, 13, 14, and 15.

Examples:

9. What will they do ?
   A. Say good bye.
   B. Change reservation.
   C. Arrive at the airport.
   D. Meet at the airport.
13. What is being advertised ?
   A. The show.
   B. The clown.
   C. The acrobat.
   D. The circus.
14. When will the show start ?
   A. First day next month.
   B. Every day.
   C. This weekend.
   D. Next week.
15. Who will get 25% discount ?
   A. The first buyer.
   B. The first ten buyers.
   C. The clowns.
   D. The family.

Discarded items:

5. Women: would you like some coffee?
   A. I like hot drinks
   B. That's not my cup
   C. No thanks, I prefer tea
7. What does Lidya do in her spare time ?
   A. She writes novels
   B. She usually buys novels
   C. She hates reading novels
   D. She likes reading novels
10. What does Winda imply ?
   A. She doesn't like Laskar Pelangi movie.
   B. She doesn't want to see that movie.
   C. She has never seen such a good movie likes Laskar Pelangi.
   D. She wants to see that movie.

The good test is a test which has good or moderate till excellent mark on the level of difficulty and discriminating power. There are 4 test items which is included in this group. The test items number are 4, 8, 11, and 12. These numbers of test items is classified as the good group.

Good items:

4. Women: what do you think about this best seller novel?
   A. It's quite interesting

The bad test is a test which has revision mark both on the level of difficulty and the discriminating power. There are 3 test items that are included in this group that should be discarded. The test items number are 5, 7, and 10. The options of the test items are bad, there are some option in the test which are not being chosen by the students. There is also option of the item which can discriminate better than the key answer. The test items cannot discriminate between the higher and the lower group too.

B. I think it's difficult to read a novel

C. I borrowed it from the library

8. What does the customer want?

A. Roasted chicken.

B. Fried chicken.

C. Boiled chicken.

D. Fresh chicken.

11. To whom is the instruction directed?

A. The guests.

B. The waiters.

C. The manager.

D. The laundry service.

12. What is expected from this instruction?

A. More customers will come to eat.

B. The guests have to prepare all things.

C. The table will be ready for the VIPs.

D. All the managers will be satisfied.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the evaluation of the test items, the researcher would like to draw some conclusions as follows. First, based on the criteria to prove the test items validity, it is concluded that the test items is *valid*. In other word, the content validity of English Listening test items for the first semester of grade twelve in SMK N 5 Pontianak in Academic Year 2013/2014 fulfill the requirement of good test items. Second, in terms of the reliability by using Kuder Richardson (KR 20), it was found out that the test is *0.183 (negligible)*. In other word, the reliability of the test items is not fulfilled the requirement of good test items. Third, the mean of items difficulty level is *0.688* which means the items classified as *moderate* test items. As the result, the whole difficulty level of the test items is fulfilled the requirements of good test items in term of difficulty level. Fourth, the mean of discriminating power is *0.256* which means the item classified as *moderate* test items. As the result, the whole discriminating power of the test items is fulfilled the requirements of good test items. And the last, there was *26* distracters that should be *revised*.

Finally, the researcher draws the conclusion that there are 4 (items number 4, 8, 11, and 12) good test items which still can be used as reference for the next summative test, 3 (items number 5, 7 and 10) test items should be discarded or changed by the other test item and 8 (items number 1, 2, 3, 6, 9, 13, 14, and 15) test items should be revised if the teacher want to use it for the next summative test.

Suggestion

Based on the conclusion above, the researcher would like to offer the following suggestions. First, it is suggested for the teacher to try out the test items and check several times to find some mistakes which may have been missed through analysis the test items related to content validity, level of difficulty, discriminating power and items distracters. Second, it is suggested for the teacher to use a good test items which are found in the results of this research and revised test items that need revision before it is used for the next summative test. Third, It is suggested for the teacher to make consistent alternatives or distracters of the test such as if it is with 4 alternatives then all of the test should with 4 alternatives, so the analysis of the test will be accurate.

REFERENCES

Best, J. W. (1977). *Research in Education.* Boston: Prentice-Hall.

Feyten, C. M. (1991). The Power of Listening Ability: An Overlooked Dimension in Language Acquisition. University of South Florida.

Fiktorius, T. (2014). A Validation Study On National English Examination Of Junior High School In Indonesia. Indonesia. Tanjungpura University.

Fulcher, G. (2010). *Practical Language Testing* . London: Hodder Education, An Hachette UK Company.

Gall, M. D. et al. (2007). Educational Research Eighth Edition. United States of America. Pearson Education, Inc.

George Morgan, P. A. (2008). *Developing Tests and Questionnaires for a National Assessment of Educational Achievement.* Washington, DC: The world bank.

Gronlund, N. E. (1977). *Constructing Achievement Test.* United States of America: Prentice Hall.

Hughes, A. (2003). *Testing for Language Teacher*. Cambridge: Cambridge University Press.

John W. Best, J. S. (2006). *Research in Education Tenth Edition.* United States of America: Pearson Education Inc.

Silver, H. (2004). Evaluation Research in Education: Press Release [online]. August 2006. Available: URL http://www.Evaluation Research in Education.htm [Accessed: 7 September 2014].

Vandergrift, L. (2004). Listening to learn or learning to listen. *Annual Review of Applied Linguistics*, 3.