ANALYSIS OF ENGLISH TRY OUT TEST OF THE NINTH GRADE STUDENTS AT MTSN PROBOLINGGO

Dea DB Cindrakasih
English Teacher of Smart Center Education Pasuruan
debteukey@gmail.com

Abstract: The specific objectives of the study are to analyze the content validity, the reliability, the item distractor, the item difficulty and the discriminating power of the English try out test of the ninth grade students at MTSN Probolinggo. The research designs are descriptive quantitative and content analysis. The subject of this study is 3 classes of the third grade students of MTSN Probolinggo with 91 students. The data was took on 9 April 2017. The researcher uses documentation and human instrument. The findings of the research show that the content validity based on the basic competence are 96% appropriate and 4% unappropriate with the syllabus. In SPSS, the content validity on each packet is low because the total numbers of valid item is under 50%. The reliability in Packet 1 is moderate and the rest packets has high criteria. Packet 1, 2, 3, 4 and 5 has 26, 35, 31, 24 and 28 options of excellent distractor. The item difficulty in Packet 1, 2, 3, 4 and 5 is there were 14, 18, 10, 11 and 16 difficult items. In the result of SPSS 23.0 of discriminating power, Packet 1, 2, 3, 4 and 5 has 13, 19, 17, 26 and 15 very good items. The conclusion is the content validity based on syllabus is high however in SPSS, the content validity is low. The reliability is moderate and high. The item difficulty is low. The item distractor is high and the discriminating power is low.

Key Words: Content validity, discriminating power, item analysis, item difficulty, item distractor and reliability.

In teaching language specifically English, teachers not only teach their students but also evaluate them. Evaluation is applied to find out students’ achievements and diagnoses students’ learning outcomes, and plays a vital role in improving the teaching quality (Jahanian, 2012: 253). In evaluating their students, there are some ways which should be done by the teacher in order to analyze students’ English skill and ability, one of them is giving a test to the students. As stated by Heaton (1998) in Primadani & Sulistyo (2013:1), both teaching and testing are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. The importance of testing as stated by Harris (1969:2-3) in Evitasari (2012:8), those are: 1) to determine readiness for instructional programs; 2) to classify or place student in appropriate language classes; 3) to diagnose the student’ strengths and weaknesses; 4) to measure aptitude for learning; 5) to measure the extent of students of achievement of the instructional goals and 6) to evaluate the effectiveness of instruction. Since that matter, it is obvious that conducting a test cannot be ignored in the implementation of teaching and learning process.

UKE says, a company registered in England and Wales, their research in 2015 stated that a test as an instrument of evaluation is a systematic procedure of description, collection and interpretation in order to measure students’ achievement ability, knowledge, and performance what they have been learned in learning process and to get a value judgment. The purpose of a test is able to give the valid information on the students’ abilities and knowledge. Test is divided into two kinds which are stated by Brown (2004:43), the first kind of tests are tests which are rarely created by teachers, language aptitude test and proficiency test, and the second kind of tests are tests which are needed to be created by teachers, placement test; diagnostic test and achievement test. Language aptitude test is a language aptitude test is designed to measure capacity or general ability to learn a foreign language and ultimate success in that undertaking. However, a proficiency test is not limited
to only any course, curriculum, or single skill in the language, but it consists of multiple-choice items on grammar, vocabulary, reading comprehension and aural comprehension. Certain proficiency test can be regarded as placement test, the aim is to place a student into a particular level in a school. In contrast with proficiency test, diagnostic test is designed to diagnose the specific aspects of a language. Achievement test which is related to classroom lessons, unit, or a curriculum. Achievement test is often regarded as summative test since it is conducted in the end of a unit or semester of teaching learning process.

To be considered as good test, there must be five mayors features, which are content validity, reliability, item distractor, item difficulty and discriminating power, in it. Sulistyo (2015:61) stated “validity is the precision of the test in measuring what it is intended to be measured”. In addition, Fulcher & Davidson (2007:6) explained that content validity is defined as any attempt to show that the content of the test is a representative sample from the domain is to be tested. It means that if the items in the test constitute a representative sample of the total course content to be tested, the test can be said to have content validity. A good test should not only be valid but also be reliable. This means that the results of a test should be dependable. They should be consistent; remain stable, should not produce different results when it is used in different days. Wordpress, a weblog for tutors and test writers of Penn State, their research in 2017 explained that reliability is similar with the consistency of a test, survey, observation, or other measuring device. Another features in a good test is there is a good item distractor which the multiple choice options that are not the correct answer. It is the features we need to analyze in a multiple choice test not in essay test. The next is item difficulty which expresses the proportion or percentage of students who answered the item correctly. In addition, discriminating power also takes role in indicating a test regarded as a good test. Oshkosh, a website of the University of Wisconsin, their research in 2012 stated that discriminating power can be calculated by ranking the students according to total score and then selecting the top 27 percent and the lowest 27 percent of the total score.

Based on the explanation above, the researcher is interested in conducting this research, which is analyzing five kinds of item on try out test in Junior High School level. This study is aimed to analyze the content validity, reliability, item distractor, item difficulty and discriminating power on English Try Out Test of the ninth grade students of MTS As-Sholchah Warungdowo-Pasuruan. The reason is because this school still uses KTSP 2006 Curriculum or Kurikulum Tingkat Satuan Pendidikan 2006, based on Keputusan Menteri Agama in 2014, general subject lessons such as English, Science, Mathemetic and etc, it needed to use KTSP 2006 Curriculum or Kurikulum Tingkat Satuan Pendidikan, however, religious lessons such as Arabic, Qurdist, Fiqih and etc, it is using 2013 Curriculum or Kurikulum 2013. In contrast with other schools in Pasuruan which use Curriculum 2013. The researcher wants to know the five items in the try out test whether it is appropriate to be tested to the students of MTS As-Sholchah or not. Another reason is because try out test is a test given to students to try the similar test with National Examination or UN . Nowadays, National Examination is being a basic evaluation which measures the competence of the students in the end of their study and is conducted every year for the third grade of junior and senior high school, also for the sixth grade of elementary school to improve the education quality in Indonesia (Fitria, 2012:1). Besides, in MTS As Scholhah, this school stil use PBT (Paper Based Test) instead of CBT (Computer Based Test) which is able to be analyzed accurately.

The researcher analyzes the try out test which consists of 50 numbers of multiple choice and is made into 5 packets, 1 packet is different with others. The reason is because this is a new educational system in making the try out test for students in Indonesia. But it does not mean every packet has different questions in it, what makes it different is only the number order.

This study focused on analyzing the content validity, reliability, item distractor, item difficulty and discriminating power on English try out test for ninth grade students at MTS As-Scholhah. The researcher analyzes the content validity in order to find out whether the test represents all the contents in the Graduate Competence Standart or SKL. Besides, she analyzes the reliability to measure the test consistency. Then she analyzes the item distractor in order to find out whether the test has a good multiple choice item in every questions and the test has good item distractor or not. The next is she analyzes the item difficulty which deals with the difficulty level in such a test. By the discriminating power, she analyzes the students level achievement in doing the try out test.
REVIEW OF RELATED LITERATURE

In teaching learning process, both teaching and testing are closely interrelated that is impossible to work in either field without being concerned with the other. Brown (2004:3-4) simply stated that “a test is a method of measuring the test takers’ ability, knowledge or performance in a given domain”. Based on this statement, he then explained that a method means a set of techniques, procedures or item which requires performance of the test takers. To qualify a test, the method must be well structured: multiple-choice test with prescribed correct answers and writing test with scoring rubric. Next, a test must the test takers’ ability, knowledge or performance in general or specific competencies. Lastly, a test measuring a given domain means in the case of language proficiency test, it may measure a general competence but it actually includes all skills of a language.

A test can categorized into written test and spoken test in the process criteria. It is categorized into the the way the tester conducting the test. Written test can be divided into two, those are: standardized test and teacher-made test. Standardized test is a test which a particular group of students will take the same test that will be scored and analyzed in the same way (Grant, 2017:1). However, teacher-made test is made by the teacher and designed to solve the problem or requirements of the class (M, 2016:1). The next test is oral test. Oral test is a test which both the questions and the answers are done orally. This test is conducted to measure the students’ speaking ability.

A test must be good enough to implemented to the test takers. Sulistyo (2015:40) explained that a good test must meet the requirements as elicitation tools or instruments, those are validity and reliability. Some experts explained some explanation about validity in language testing and assessment.

Validity is one of some requirements of a good test. A good test must be valid. Brown (2001:) simply explained validity is the degree of the test to which the actually measure what intended to measure. However, Arikunto (2003:65) stated that validity in testing and assessment does not only mean discovering whether a test measures accurately what is intended to measure but it deals with the results of the test, in other words, it deals with the score.

However, reliability refers to the consistency of students’ scores would be received on alternate forms of the same test (Wells & Wollack, 2003:2). It means that reliable test can produce stable or consistent score no matter who administrates the test. The scores are also consistent no matter when and where the test is administered.

Besides, there is another way to examine the quality of individual items in a test, especially in multiple choice test type, that is item analysis. Through item analysis, the items quality can be improved in the terms of its form and internal attributes. Besides, item analysis is necessary for increasing test makers’ skills in constructing the test items, and identifying the course content which need to be emphasized. Alfaraby (2009:1) stated that the result of item analysis can be used to select the best item difficulty that best to discriminate the high and low students. According to Sulistyo (2015:209), classical item analysis is mainly aimed at three aspects, those are: item difficulty, discriminating power and item distractor in terms of multiple choice items.

Item distractor deals only with multiple choice items which is concerned with the effectiveness of distractors utilized in the option. Sulistyo (2015:223) states that a multiple choice item consists of a stem and options or alternatives following the stem. The options provide a correct answer or key and distractors.

Then item difficulty simply defined as the proportion or percentage of students who answered the item correctly. In the other words, item difficulty may be understood as the comparison between the numbers of the test takers who anwser the items correctly with the total numbers of the test takers answering to that item (Sulistyo, 2015:211). The comparison is commonly called as the proportion which is expressed using an index ‘p’.

An item discrimination is aimed to examine whether an item can discriminate the test takers who possess the competence from those who do not (Sulistyo, 2015:216). In addition, according to Washington Edu (2017:1), item discrimination refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested. The point-biserial correlation is an index of item discrimination, which is written as D.

There were some other researcher findings which have similar analysis on English Test. In this research, the researcher focuses on analyzing English Try Out Test given to ninth grade students at MTSN Probolinggo. She also analyses five items in the test, which are content validity, reliability, item distractor, item difficulty and discriminating power. In the research by Indrawati (2016) entitled “Item Analysis on Try Out Test in Third Grade of SMPN 10
Pasuruan Academic Year 2015-2016”, it is found that the researcher applied this research in SMPN 10 Pasuruan because she wanted to know how the quality of the try out test item which was made by the teachers. The subject of the research was a half of the students in third grade which amount is 50 students. After analysing the data, she found the following results: the test had 43 or 86% items appropriate with the syllabus and 7 or 14% items not appropriate with the syllabus. Based on SPSS 23.0, the content validity was low; there were 18 valid items or 26% of the total items and 34 invalid items or 64% of the total items. The test had high reliability because it found that it was 0.724 by using Cronbach Alpha, the \( r \) table was more than 0.279. The item difficulty was moderate by seeing its Mean because there were 15 easy items and 9 difficult items. The discriminating power was categorized into poor criteria because there were 12 good items, 6 items need to be received and revised, 7 items need to revised and 25 items rejected. It means that a half of the items was rejected. Based on the result, the researcher concluded that test was average. It can be used but some items need to be revised. The test was low in content validity, high in reliability, moderate in item difficulty and poor in discriminating power.

**METHOD**

In this research, the researcher uses two kinds of design, those are: descriptive quantitative and content analysis. The object of the study is English try out test 2016/2017 in the form of multiple choice which is published by Ministry of Religion and designed by a group of English testmaker who cooperate in Ministry of Religion. There are 5 sets of English try out test and each packet has different items. There were 18 students who did the packet 1, 19 students who did the packet 2, 18 students who did the packet 3, 19 students who did the packet 4 and 17 students who did the packet 5. The researcher analyzes the English try out from Ministry of Religion (KEMENAG) because the test will determine if the students pass the degree of test and give some feedback to the students in National Examination test.

In this research, the researcher uses documentation to get some information and human instrument for the primary instruments of this research. The test was held on 9 March 2017. Meanwhile, the researcher took the data on 9 April 2017.

In analyzing the data, the researcher uses some steps, they are: editing, coding, tabulating and analyzing the data using SPSS. In analysing using SPSS, the researcher uses the simple formula of Spearman Rank technique, the correlation value (\( r \)) between two variables.

### Step to Determine the Content Validity of the Test

The test consists of 50 items, in the form of multiple choice. The researcher tries to analyze the content validity based on standart competence of National Examination in Indonesia which is used by Kemenag. The following table is used to check the content validity based on the standart competence:

<table>
<thead>
<tr>
<th>No.</th>
<th>Basic Competence</th>
<th>Related Items</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to Widiyanto (2010:34-37), correlation coefficient uses the Pearson Formula in SPSS:

\[
r_{xy} = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2}}
\]

Notes :
- \( r_{xy} \) = correlation coefficient
- \( X \) = total score
- \( x \) = item score
- \( n \) = number of subject

### Step to Determine the Reliability of the Test

The reliability of measuring instruments is the degree of consistency with measure. In this research, the researcher measured the reliability of each item using Cronbach’s Alpha in SPSS, the formula is as follows:

\[
a = \left( \frac{k}{k-1} \right) \left[ \frac{sr^2 - \sum_1^k s^2_{i}}{sX^2} \right]
\]
Step to Determine Item Distractor

Distractors is used to distract the test takers with a low level of mastery by giving nearly the similar answer with the correct answer. To examining the item distractor, the researcher used the formula, as follows:

\[
IP = \frac{P}{(N - B)(n - 1)} \times 100\%
\]

Notes:
- IP = distractor index
- P = the total number of test takers who answer the item distractor
- N = the total number of the test takers
- B = the total number of the test takers who respond to the correct answer
- n = the total number of the optional answers
- l = number

The result on the reliability will be categorized into 4 categories, as follows:

<table>
<thead>
<tr>
<th>Degree of Reliability</th>
<th>Item Test Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 0.90</td>
<td>Perfect</td>
</tr>
<tr>
<td>0.70 – 0.90</td>
<td>High</td>
</tr>
<tr>
<td>0.50 – 0.70</td>
<td>Moderate</td>
</tr>
<tr>
<td>Less than 0.50</td>
<td>Lower</td>
</tr>
</tbody>
</table>

**Table 2 The Degree of Reliability Criteria**

Step to Determine Discriminating Power

The researcher divides the students into two groups, that are upper group and lower group. In identifying these two groups, discriminating power allows the researcher to distinguish the performance of the students in upper group with the students in lower group. The discriminating power must be interpreted in the rank scale of discriminating power by using the formula, as follows:

\[
D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B
\]

Notes:
- J = the total number of the test takers
- J_A = the total number of the upper group
- J_B = the total number of the lower group
- B_A = the total number of the upper group which responds the item correctly.
- B_B = the total number of the lower group which responds the items correctly.
- P_A = the proportion of the total number of the upper group which responds the item correctly.
- P_B = the proportion of the total number of the lower group which responds the item correctly.

In Statistical Packed for Social Science (SPSS), the way to measure the items is to see the value of Pearson Correlation by using r table. Then, the researcher divided it into two groups, upper group and lower group using median, so the division is 27% of the upper group and 27% of th lower group. The tables for upper and lower group, are as follows:

**Table 4 Criteria to Interpret Item Difficulty**

<table>
<thead>
<tr>
<th>Degree of Item Difficulty</th>
<th>Quality of Item Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 0,70</td>
<td>Easy</td>
</tr>
<tr>
<td>0,30 – 0,70</td>
<td>Moderate</td>
</tr>
<tr>
<td>&lt; 0,30</td>
<td>Difficult</td>
</tr>
</tbody>
</table>

Step to Determine Item Difficulty

Item difficulty also needs to be calculated in SPSS to find out difficulty level of the test.in SPSS, the researcher used the value that is shown as MEAN in statistic table to extend the difficulty. the formula used to examine the proportion correct (p) is, as follows:

\[
p = \frac{\sum B}{N}
\]

To interpret the item difficulty, the researcher uses the criteria as shown in the table below, as follows:

**Finding of the Content Validity**

Based on table 4, it shows that there are 50 items in each packet. Each packet has the same ba-
sic competence but different items in English Try Out test for ninth grade. There are 4 items (8% of 50 items) in the first basic competence, 14 items (28% of 50 items) in the second basic competence, and 30 items (60% of 50 items) in the third competence. There are 2 items are not invalid in the basic competence, those are are in number 1 and 2. The result of content validity of Try Out Test is high because the total valid items is 96%. Besides, in the following table will be shown the content validity on each packet that is calculated using SPSS 23.0.

The validity can be seen from the total significant. According to Santoso (2016:358), if the significant (2 tailed) < 0.025, the item is valid. To determine the value of \( r \), the researcher use \( df = N-2 \). To determine the \( r \) value, the researcher calculate it based on Pearson Product-Moment Correlation table. the \( N \), \( r \) and the \( df \) on each packet will be shown in the table below.

### Table 5 The Interpretation of the total number of students, \( r \) and the \( df \).

<table>
<thead>
<tr>
<th>Packet</th>
<th>( N )</th>
<th>( Df )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>16</td>
<td>0.497</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>17</td>
<td>0.482</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>16</td>
<td>0.497</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>17</td>
<td>0.482</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>15</td>
<td>0.514</td>
</tr>
</tbody>
</table>

The item is categorized as valid items if the item’s validity is more than the \( r \) value. And, an invalid item means its validity is less than the \( r \) table. Here are the result fo the content validity in all packets, as follow:

### Table 6 The Content Validity calculated by SPSS

<table>
<thead>
<tr>
<th>Packet</th>
<th>Valid</th>
<th>Invalid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14%</td>
<td>86%</td>
</tr>
<tr>
<td>2</td>
<td>28%</td>
<td>72%</td>
</tr>
<tr>
<td>3</td>
<td>36%</td>
<td>78%</td>
</tr>
<tr>
<td>4</td>
<td>34%</td>
<td>66%</td>
</tr>
<tr>
<td>5</td>
<td>18%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Based on the table above, it can be concluded that all packets has low validity because the total number of valid items is below 50%.

In conclusion, the validity of English Try Out of ninth grade calculated with SPSS 23.0, the result is, there are 14% or 7 valid items in packet 1, 28% or 14 valid items in packet 2, 22% or 11 valid items in packet 3, 34% or 17 valid items in packet 4 and 18% or 9 valid items in packet 5.

### Finding of Reliability

In the research method, the researcher explained about Cronbach Alpha for measuring reliability. Here is Cronbach’s Alpha of English Try Out Test and its category in all packets, as follows.

### Table 7 Reliability Cronbach’a Alpha

<table>
<thead>
<tr>
<th>Packet</th>
<th>Cronbach’s alpha</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.636</td>
<td>Moderate</td>
</tr>
<tr>
<td>2</td>
<td>0.705</td>
<td>High</td>
</tr>
<tr>
<td>3</td>
<td>0.768</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>0.878</td>
<td>High</td>
</tr>
<tr>
<td>5</td>
<td>0.747</td>
<td>High</td>
</tr>
</tbody>
</table>

In short, the results is all packet is reliable. The reliability of English Try Out Test for ninth grade in each packet are moderate in packet 1, high in packet 2, high in packet 3, high in packet 4 and high in packet 5.

### Finding of Item Distractor

To measure the distractors quality of each item, the researcher used the formula as follows:

\[
IP = \frac{x}{100}\%
\]

The calculation can be seen in Appendix 3. The following table is the classification of the item distractor of the try out test in Packet 1, as follow:

### Table 8 The Interpretation of Item Distractors

<table>
<thead>
<tr>
<th>Packet</th>
<th>Excellent</th>
<th>Good</th>
<th>Deficient</th>
<th>Poor</th>
<th>Very poor</th>
<th>Unidentified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>37</td>
<td>29</td>
<td>44</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>33</td>
<td>14</td>
<td>48</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>30</td>
<td>27</td>
<td>43</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>33</td>
<td>24</td>
<td>43</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>27</td>
<td>16</td>
<td>49</td>
<td>24</td>
<td>8</td>
</tr>
</tbody>
</table>
From the above, in English Try Out Test Packet 1, there are 26 options of excellent distractor, 37 options of good distractor, 29 options of deficient distractor, 44 options of poor distractor and 14 options of very poor distractor. However, in English Try Out Test Packet 2, there are 35 options of excellent distractor, 33 options of good distractor, 14 options of deficient distractor, 48 options of poor distractor and 20 options of very poor distractor. Then, the result of the item distractor in English Try Out Test Packet 3 is there are 4 options cannot be identified. Other results are found that there are 31 options of excellent distractor, 30 options of good distractor, 27 options of deficient distractor, 43 options of poor distractor and 24 options of very poor distractor. In English Try Out Test Packet 4, there are 4 options cannot be identified as found in Packet 3. In this packet, it is also found that there are 24 options of excellent distractor, 24 options of good distractor, 24 options of deficient distractor, 43 options of poor distractor and 24 options of very poor distractor. The last in Packet 5, the result of item distractor classification in English Try Out Test Packet 5 as shown in the table above, there are 28 options of excellent distractor, 27 options of good distractor, 16 options of deficient distractor, 49 options of poor distractor, 24 options of very poor distractor and 8 options cannot be identified.

**Finding of Item Difficulty**

Item difficulty is the difficulty level of each item. Good item has balance in difficulty level (proportional), so it means the item is not too difficult or too easy. Based on the SPSS output table, the value in MEAN is to extent the difficulty level. The following table is the result of item difficulty in all packets.

<table>
<thead>
<tr>
<th>Packet</th>
<th>Easy</th>
<th>Moderate</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

**Table 9 Interpreting of Item Difficulty**

In conclusion, in Packet 1, there were 13 very good items, 7 reasonably good items, 4 marginal items and 26 poor items. Packet 2 has r in level significant 5% = 0.456 which based on the correlation table product moment. The result is that there are 19 very good items, 4 reasonably good items, 6 marginal items and 20 poor items. In Packet 3, based on the correlation table product moment in level significant 5% = 0.468, it is found that Packet 3 has 17 very good items, 7 reasonably good items, 5 marginal items and 20 poor items. Then, in Packet 4, based on the correlation table product moment in level significant 5% = 0.456. It is found that there are 26 very good items, 6 reasonably good items, 7 marginal items and 11 poor items. The last is in Packet 5, the r in level significant 5% = 0.482. So, it is found that this packet has 15 very good items, 9 reasonably good items, 6 marginal items and 20 poor items.

**Finding of Discriminating Power**

To measure the discriminating power, the researcher split the data into upper group and lower group. This study focuses on five packet for English Try Out test, so the researcher split the data in each packet. The way to split the data is by seeking the 27% students from upper group and lower group. There are 5 students each in upper group and lower group in all packets. Here are the interpreting of discriminating power and its category in all packets.

**Table 10 Interpreting of discriminating power**

<table>
<thead>
<tr>
<th>Packet</th>
<th>Very good item</th>
<th>Marginal item</th>
<th>Reasonably good</th>
<th>Poor item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>7</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>4</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>7</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>6</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>9</td>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>

In packet 1, it is found that there are 13 very good items, 7 reasonably good items, 4 marginal items and 26 poor items. Packet 2 has r in level significant 5% = 0.456 which based on the correlation table product moment. The result is that there are 19 very good items, 4 reasonably good items, 6 marginal items and 20 poor items. In Packet 3, based on the correlation table product moment in level significant 5% = 0.468, it is found that Packet 3 has 17 very good items, 7 reasonably good items, 5 marginal items and 20 poor items. Then, in Packet 4, based on the correlation table product moment in level significant 5% = 0.456. It is found that there are 26 very good items, 6 reasonably good items, 7 marginal items and 11 poor items. The last is in Packet 5, the r in level significant 5% = 0.482. So, it is found that this packet has 15 very good items, 9 reasonably good items, 6 marginal items and 20 poor items.
DISCUSSION

Content Validity

Content validity is the degree in which the test is actually measured what is intended to be measured. According to Brown (2001:388), content validity is the extent to which a test actually samples the subject matter and if the test takers perform the behaviour what is being measured.

The researcher has already analyzed the Try Out Test at MTSn Probolinggo. The basic competences are divided into three levels, those are, Knowledge and Understanding (Pengetahuan dan Pemahaman), Application (Aplikasi), and Reasoning (Penalaran), all includes some speaking and writing texts such as, transactional interaction/written interpersonal (apologizing, wish, hope), short functional texts (announcement, notice, label), descriptive, recount, narrative, procedure and report. There are 5 item packets, which have the same basic competence on each item. The researcher found that there are 4 items (8% of 50 items) related to the material in Knowledge and Understanding level, 14 items (28% of 50 items) in Application Level and 30 items (60% of 50 items) in Reasoning Level. There are 2 items are not invalid in the basic competence, those are are in number 1 and 2.

The researcher also used SPSS 23.0 to calculate the content validity or not by significant (2 tailed) < 0.025, the item is valid. In Packet 1, there are 18 test takers, so, based on Pearson Product-Moment Correlation table, the r = 0.497. It means that if the significant probability > r, then the item is valid. The result is there are 14% or 7 valid items and 86% or 43 invalid items. However, the result found in Packet 2, there are 28% or 14 valid items and 72% or 36 invalid items. This packet was taken by 19 students. So, based on Pearson Product-Moment Correlation table, the r = 0.482. Packet 3 was taken by 18 students. So, based on Pearson Product-Moment Correlation table, the r = 0.497. The valid items are 22% or 11 and the invalid items are 78% or 39. Then, the result in Packet 4 with the r = 0.482 is, there are 34% or 17 valid items and 66% or 33 invalid items. This packet has 19 test takers. The last in Packet 5, the r = 0.514 with 17 test takers. The results of valid items and invalid items are 18% or 9 and 82% or 41.

The result of content validity of Try Out Test based on the basic competence is high because the total valid items is 48 items or 96% and The invalid items are only 2 or 4% items. Related to the previous study by Indrawati (2016:1), in her study, the test had 43 or 86% items appropriate with the syllabus and 7 or 14% items are not appropriate with the syllabus. And based on SPSS 23.0, there were 18 valid items or 26% of the total items and 34 invalid items or 68% of the total items.

Reliability

Reliable test produces stable or consistent result, no matter where and when the test is administered. It also refers to the consistency of the students’ scores would be received on alternate forms of the same test. According to Sulistyo (2015:43), a reliable test should be demonstrated by the score consistency within raters or interraters, between raters and interraters and across time and place. If the test is taken by different students then the results is consistent, it means the test has high reliability.

Reliability means the stability of the score. The researcher used SPSS 23.0 to know the reliability of the test. The degree of reliability is if Alpha > r table = consistent or reliable but if Alpha < r table = inconsistent or unreliable. The result in Packet 1, the reliability of Cronbach’s alpha is 0.635. The compare of r table is 0.468 and the Cronbach’s alpha is 0.636. It means that it is reliable and the reliability is in moderate criteria.

Then, Packet 2 has a reliable and high reliability with Cronbach’s alpha value is 0.705. In Packet 3, the reliability of Cronbach’s alpha is 0.768. It means that it has a reliable and high reliability. However, in Packet 4, the compare of r table is 0.468 and the Cronbach’s alpha is 0.878. So, it is also categorized as reliable and high reliability. The last is in Packet 5, the reliability of Cronbach’s alpha is 0.878, so Packet 5 is reliable and its reliability is in high criteria.

Related to the previous study done by Indrawati (2016:1), the test had high reliability because it is found that the Cronbach’s Alpha is 0.724, with the r table > 0.279.

Item Distractor

Item distractor deals with the multiple choice options which is concerned with the effectiveness of distractors utilized in the options. According to Sulistyo (2015:224). Distractor is used to distract the low level test takers by giving nearly the similar answer with the correct answer. Arifin (2009:279) stated that a good item, its distractors will be choosen evenly by the students. On the contrary, a poor item, its distractors will be choosen randomly by the students.

The criteria of the item distractor divided into excellent, good, deficient, poor, and poor items. Excellent, good and deficient items happen if several
students choose the distractor. However, poor item can be happened because no students choose the distractors. There are also some item cannot be identified because all students choose the correct answer, not the other distractors.

The researcher found that each packet of the test has different item distractor category. In packet 1, there were 26 options of excellent distractor, 37 options of good distractor, 29 options of deficient distractor, 44 options of poor distractor and 14 options of very poor distractor. Then, in packet 2, there were 35 excellent distractor options, 33 good distractor options, 14 deficient distractor options, 48 poor distractor options and 20 very poor distractor options. However, in packet 3, there are 4 options cannot be identified, 31 options of excellent distractor, 30 options of good distractor, 27 options of deficient distractor, 43 options of poor distractor and 14 options of very poor distractor. In packet 4, there are 4 options cannot be identified as found in packet 3. In this packet, it is also found that there are 24 options of excellent distractor, 33 options of good distractor, 24 options of deficient distractor, 43 options of poor distractor and 24 options of very poor distractor. The last in packet 5, there were 28 options of excellent distractor, 27 options of good distractor, 16 options of deficient distractor, 49 options of poor distractor, 24 options of very poor distractor and 8 options cannot be identified. Related to the previous study done by Indrawati (2016:1), the test had no excellent item and only 1 good item. There were also 25 deficient items, 7 poor items and 17 very poor items.

**Item Difficulty**

Item difficulty is the proportion or percentage of students who answered the item correctly. Difficulty index illustrates how easy and difficult an item. According to Arifin (2009:266), if an item has the proportional difficulty index, it can be said that the item is not good enough. A good item should not be too difficult and not too easy. An easy item cannot motivate the students to heighten their ability to solve the questions. In other words, a difficult item distracts students to not solve the questions based on their ability because students may think that the item is too hard.

The classification of the difficulty index can be divided into 3 categories, those are: if the item is < 0.30, then the item is difficult; if the item is between 0.30-0.70, then the item is moderate; if the item is > 0.70, then the item is easy. Based on the table in SPSS 23.0, by looking at the MEAN of each item, the researcher found that each item's difficulty level on each packet.

In packet 1, there were 16% or 8 easy items, 56% or 28 moderate items and 28% or 14 difficult items. Then, in packet 2, there are 24% or 12 easy items, 40% or 20 items and 36% or 18 difficult items. Meanwhile in packet 3, there are 24% or 12 easy items, 56% or 28 moderate items and 20% or 10 items difficult items. In packet 4, there are 32% or 16 easy items, 46% or 23 moderate items and 22% or 11 difficult items. The last in packet 5, there are 36% or 18 easy items, 32% or 16 moderate items, 32% or 16 difficult items.

**Discriminating Power**

Discriminating power is aimed to differentiate between the students with good ability and the students with lack ability. According to Sulistyo (2015:216), an item discrimination is to examine whether an item can discriminate the test takers who possesses the competence from those who do not. Arifin (2009:273) stated that high coefficient of the item discriminating index means an item can discriminate the students who have the ability from those who do not. The discriminating power identifies the upper and lower group of the students. In this study, the researcher found that on 5 packets, there were 5 students in upper group and 5 students in lower group on each packet. So, each packet has the same total number of upper and lower group.

Besides, the researcher used the formula to calculate the discriminating power by using SPSS 23.0 and seeking the result by the table Pearson Correlation. In packet 1, the result of Pearson Correlation compare with the t table with the total data = 10. Based on the correlation table product moment in level significant 5% = 0.468. It is found that there are 13 very good items, 7 reasonably good items, 4 marginal items and 26 poor items. Packet 2 19 very good items, 4 reasonably good items, 6 marginal items and 20 poor items. In Packet 3, based on the correlation table product moment in level significant 5% = 0.468, it is found that Packet 3 has 17 very good items, 7 reasonably good items, 5 marginal items and 20 poor items. Then, in Packet 4, it is found that there are 26 very good items, 6 reasonably good items, 7 marginal items and 11 poor items. The last is in Packet 5, it is found that this packet has 15 very good items, 9 reasonably good items, 6 marginal items and 20 poor items.
Related to the previous study by Indrawati (2016:1), in her study, it was found that the discriminating was in poor criteria, there were 12 good items, 6 items need to be received and revised, 7 items need to be revised and 25 items rejected.

**CONCLUSION**

Based on findings and discussion, it can be concluded that this English Try Out test has 96% or 48 items appropriate with the basic competence and 4% or 2 items not appropriate with the basic competence in the content validity. On the other hand, in examining the content validity using SPSS 23.0, the researcher analyzed the content validity on each packet. The result is Packet 1, 2, 3, 4 and 5 has 14% or 7, 28% or 14 valid items, 22% or 11, 34% or 18% or 9 and 82% or 41 for valid items.

The reliability in Packet 1 is in moderate criteria and for the rest packets, that is in Packet 2, 3, 4 and 5 is in high criteria.

The item distractor was calculated by the formula. In Packet 1, 2, 3, 4 and 5 has 26, 35, 31, 24 and 28 options of excellent distractor. In calculating the item difficulty of English Try Out Test, the researcher used SPSS 23.0 and identified it by MEAN. The result is Packet 1 has 14, 18, 10, 11 and 16 difficult items.

In discriminating power, the researcher divided the test takers into upper and lower group then analyzed it using SPSS 23.0. All packets have the same total number of the students or the test takers both in upper and lower group, that is 5 students. In SPSS 23.0, there were 13, 19, 17, 26 and 15 very good items in Packet 1, 2, r, 4 and 5.

Finally, the researcher will give some suggestion based on the result of the study. The suggestion is for the test makers and the other researchers. The test makers should create an English Try Out test that has appropriate content with the syllabus. Besides, in creating the try out test, the test makers should consider to make good distractors to distract the students well in order to avoid identifications options in an item. The other researchers are expected to use this study as references to analyze the content validity, reliability, item distractor, item difficulty and distracting power in an English test.

**REFERENCES**


Evitasari, Aprilia.2012. *Quality of Test for the Summative Test at the Second Year Students of Mts Al-Ihsan Tidu-Pohjentrek*. Unpublished S1’s Thesis. Pasuruan: Department od Education of STKIP PGRI Pasuruan


Primadani, E. A & Sulistyon, G. H. 2013. *An Analysis of a Midterm English Test of the Seventh Grade Accelerated Class at SMPN 3 Malang*. State University of Malang


