ANALYZING THE QUALITY OF FIRST ENGLISH SIMULATION TEST FOR NINTH GRADERS IN ACADEMIC YEAR 2016/2017

Muhammad Wahyuda Erdyan Mulyono, Ninik Suryatiningsih, N. Hari Gede
MTs. Negeri Pasuruan
YudaRdyan2010@gmail.com

Abstract: This research was used to know the quality of first English simulation test for ninth graders in academic year 2016/2017 in SMPN 7 Pasuruan. The quality consists of validity, reliability, item difficulty, discriminating power, and item distractors. This research was designed into descriptive quantitative which produced the result by expressing numbers. The subject of this study were 258 ninth grade learners. The data was collected on 17th February 2017 using documentation. The result of this research was Code 1 had 96% as good validity, had 0.67 as high reliability, had 52% moderate items of item difficulty, had more satisfactory items of discriminating power (26%), had 71% unworkable distractors in upper group and 57% in lower group, and had 21% inappropriate distractors. Code 2 had 100% as good validity, had 0.64 as high reliability, had 44% moderate items of item difficulty, had more satisfactory items of discriminating power (46%), had 72% unworkable distractors in upper group and 45% in lower group, and had 17% inappropriate distractors. Code 3 had 100% as good validity, had 0.37 as low reliability, had 48% moderate items of item difficulty, had more satisfactory items of discriminating power (38%), had 45% unworkable distractors in upper group and 30% in lower group, and had 26% inappropriate distractors. In conclusion, the researcher stated that each code had one lack to fulfill characteristics as a good test. Code 1 and Code 2 could not fulfill workable distractors. In other hand, Code 3 did not have high reliability.

Keywords: discriminating power, item difficulty, item distractors, reliability, validity.

Good test must have some characteristics which represent whether the test is appropriate to be given or not, and whether the test is effective to measure the achievement or not. Test can be considered as good test if it has two essential characteristics, those are, reliability and validity (Arifin, 2011). Another expert supports this statement by saying three characteristics, those are, item difficulty, discriminating power and item distractors that must be had by good test as deeper parts (Weiss, 2005). Based on these, how to decide whether one test is perceived as good test or not ought to be related to those characteristics. Validity, a good test needs to be valid so that the test is able to be in exact intention of what to be measured. According to Roa (2009), a test is considered as having a valid when it is applied in exact purpose. So, validity is the state of being legally acceptable in order to measure something appropriately. Then, reliability is created by “reliable” word. It means stable or consistent. The test must have reliability means it can produce the stable score or consistent score (Sulistyo, 2015). The stable score or consistent score is capable of being gained by the learners’ response of the test and it does not matter who does the test or when the test is administered. In item difficulty, it talks about the deep part of the test. In every test there must be the items or a test must be arranged by its items. Those items represent their independence of level in difficulty. Test must not have too difficult or too easy items. If the test has them, it can not be called as a good test and supposed to be revised. Discriminating power is used to know the classification between the learners’ good achievement and bad achievement. Related to Heaton (2010) reveals that by having discriminating power, a test can give the information about the range between the learners who understand and the learners who lack of what test demand. Item distractors have the use as decoy in each item of the test. Item distractors are settled as unavoidable part of the test in order to influence learners’ performance. Distractors infrequently influence the learners who understand enough about the material and decrease the possibility to select the distractors whereas for
the learners who do not acquire the material, distractors will be confusing option to decide the right answer. Based on Badan Standard National Pendidikan (2006) stated that mostly the learners who choose the distractors are from the lower group. In the meantime, the appropriateness of item distractors should be in use for cheating the learners who don’t master the test.

Those characteristics for deciding whether a test is good or not can be analyzed by using research to analyze the quality. Analyzing test quality is also the way to see irrelevant aspects contained. It can be understood as the important step of constructing a test and it can examine the quality of each individual item in a test. Why the parts of test needs examining because each of them characterizes the quality of a whole test. Well qualified test will reveal the test’s result accurately. It shows that analyzing the quality is worth accomplishing.

Based on the explanation above, the researcher would like to concern on conducting this research to analyze the quality of first English simulation test. The researcher focused on the first simulation test because he thought it would be the basic of arranging the next test. In fact, there would be some simulation tests of national examination so that the first simulation test was needed to make sure having good quality. This test included as teacher-made test which means its quality has more possibility to be ignored by the maker than other standardized-test. This simulation test has been formatted by the teacher as test maker into 3 different codes; code 1, 2 and 3 of the test with multiple choices with 4 alternatives; A,B,C,D. The items of each question in each code were surely different. Simulation test has main functions to be a judgment to decide next method and to know what materials which need to be emphasized in teaching and learning process. So, it must have good validity, have high reliability, have no too difficult or too easy item of item difficulty, have good discriminating power, and have effective and appropriate item distractors. In this case, the researcher conducted this test analysis to the first English simulation test to know whether the test was valid based on all material in lattice or not, to know whether the test could produce consistent score result or not, to know whether there was no too difficult or too easy item or not, to know the discriminating power to the high and low achievement of the learners, and to know the effectiveness and appropriateness of item distractors. By doing this research, the researcher would like to present test’s quality. The researcher chose this school because the researcher had an experience to teach there and he knew well the way the teacher how to apply the educational system there. Having said this, the researcher had a willingness to conduct this research of analyzing the quality of first English simulation test for ninth graders in SMPN 7 Pasuruan in academic year 2016/2017 in order to show the importance of understanding the quality of the test which is appropriate to be administered to the learner in measuring the achievement.

Accentually, there were other researches that had the same aim and willingness to this researcher to deal with measuring quality of test. This researcher investigated and discussed those other researches as the follow; Ali (2015) conducted the same research of analyzing the quality of test with the title “Test Quality Analysis of English Try Out Test For National Final Examination of Madrasah Aliyah (MA) An-Nur Winongan”. This research focused on analyzing characteristics which represented test’s quality. It showed the result with 0.55 degree of validity which means in less good category, 0.32 coefficient as low reliability, item difficulty was in easy level with 22 items, discriminating power was dominated by poor category with 31 items and deficient item distractors. Second, Indrawati (2016) did the research to know the quality of try out test. The title of her research was “Item Analysis on Try Out Test in Third Grade of SMPN 10 Pasuruan Academi Year 2015-2016”. The researcher in this study revealed the result that she got by informing that validity presented 52% as less good category, high reliability with 0.724 coefficient, moderate item difficulty because of 26 items, and 25 items of poor discriminating power with none of good category. In conclusion, from those criterions to be a good test, only reliability and item difficulty which fulfilled the characteristics as a good test. Having said these, the researcher in this study had more willingness to analyze the quality of test. Beside, looking at important role of test in teaching and learning process, but also only by administrating test learners’ abstract, that is, competence or ability can be elicited. This research wanted to focus on the same test by presenting the quality of first English simulation test for ninth graders in SMPN7 Pasuruan academic year 2016/2017 but in little different way. In one of the researches above had same focus on measuring workable item distractors but he didn’t arrange tabulation as real evidence. In this research, arranging tabulation in deciding item distractors was highlighted. The researcher not only measured workable or effectiveness but also the appropriateness of item distractors. Another difference, both research that measured the validity of try out
got less good category in matching the material with syllabus. In this research, focused on the same object, that was, try out or simulation test but the researcher would like to check validity in matching with grille or lattice not syllabus. The other differences were adding diagram and chart in measuring reliability, item difficulty, and discriminating power to show the skew and percentage.

REVIEWS OF LITERATURE

A test is not something that is done in a careless or haphazard manner. There are some characteristics that are observed and analyzed to decide the goodness of test. And these characteristics are desirable for all tests. Some of these are discussed under the various headings, include:

Validity

A good test should be valid. Validity of tests means that a test measures what it is supposed to measure, or a test is suitable for the purposes for which it is intended. That’s why validity is understood as the most important characteristic of a good examination (Fulcher & Davidson, 2007). It is vital for a test to be valid in order to interpret and apply the result accurately. According to Lukke (2014) stated that there are some factors related to test-maker which affect the result of validity, those are, (a) Test maker’s experience shows that the more his/her has an experience in creating test the high validity accepted, (b) Educational level shows test maker’s good quality will produce good test, (c) Training on constructing and analysis is needed to further enhance test maker’s skill, (d) Ability in inputing what a test needs proportionally. Test can be said to have a real high validity if the spreading the proportional accentually highlighted, (e) Then, length of the test.

Reliability

A good test should be reliable. Reliability simply means measuring what it purports to measure consistently (Brown, 2000). On a reliable test, test maker can be confident that someone will get more or less the same score on different occasions or when it is used by different people. Fulcher & Davidson (2007) said that whenever a test is administered, the reliability gives the assurance that the results could be replicated if the same individuals are tested again under similar circumstances. This desired consistency (or reproducibility) of test scores is called reliability. In other words, it is assumed that a score ought to be meaningful and interpretable, the sum of the parts should be reproducible. Indeed, the result of the test should be approximately the same. The following factors will influence the reliability of a test (Osman, 2010):

1. Administration. It is essential that each test taker has the same time, equipment, assistance, and examination environment. Test directions should be strictly enforced.
2. Scoring. Objectivity in scoring contributes to reliability. Every effort should be made to obtain uniformity of scoring standards and practices.
3. Standards. The standards of performance that are established for one class should be consistent with those used in other classes. A change in grading policies not based upon facts, uniform standards, and experience factors gained from other classes will affect the reliability of test results.
4. Instruction. The reliability of test result will be affected if the instruction presented to a class tends to overemphasize the teaching points included in the examination. This is often known as “teaching the test” and is undesirable. When the instructor gives test takers obvious clues as to the test requirements, instructor not only affects the reliability of the test, but instructor insults the intelligence of the class.
5. Length. The more responses required of test takers, the more reliable will be the test or measuring device.

Measuring reliability can be done in patterns of the Pearson Product Moment and Spearman Rank Order (Nunan, 2004). But deciding the formula should be based on data and way of estimate reliability. Common useful pattern to measure reliability in internal consistency is using K-R20 or K-R21 by Kuder Richardson (Sulistyo, 2015).

\[
K-R \ 20 \ \rho_{KR20} = \frac{k}{k-1} \left( 1 - \frac{\sum_{j=1}^{k} P_j q_j}{\sigma^2} \right)
\]

\[
K-R \ 21 \ \rho_{KR21} = \frac{k}{k-1} \left[ 1 - \frac{\mu (k - \mu)}{k \sigma^2} \right]
\]

Item Difficulty

According to Kubizyn & Borich (2007), item difficulty is the percentage of learners who respond each question of the test correctly. It is commonly determined from the proportion of the total group selecting the correct answers. In deciding item diffi-
culty, it is assumed that items of the test should not be too easy or too difficult for test takers. This is supported by the theory from Purwanto (2009), a good test item should have moderate difficult level. There are 3 criterions of item difficulty, those are, Difficult ($P \leq 0.30$), Moderate ($0.30 < P \leq 0.70$), and Easy ($P > 0.70$) with formula by Dubios. Arikunto (2008) stated that the item outside of this proportion ($0.1 < P < 0.9$) must be changed because it is considered as too easy or too difficult test.

Dubios’ Formula;

$$P = \frac{E}{Je}$$

According to Katz et al (2002) said that there are causes that effect item difficulty, those are, the item may be miskeyed, the ambiguous item or not written clearly, the item may have more than one correct answer, and the item may be too challenging relative to overall level of participants’ ability. If the item indicates too difficult or too easy based on the test result, this means the item need to be revised to bring the betterment of its difficulty.

**Discriminating Power**

Discriminating power assumes that the responses to individual items are capable of discriminating between higher ability and lower ability test takers (Heaton, 2010). Katz et al (2002) said that discriminating power is greatly influenced by item difficulty. Based on Djiwandono (2004) there is a formula with five criterions to interpret the result of discriminating power, such as Good ($D > 0.50$), Satisfactory ($0.20 < D < 0.50$), Poor ($D < 0.20$), No discrimination power ($D = 0$) and Negative ($D = -(negative)$).

Formula in deciding discriminating power;

$$D = \frac{BA}{JA} - \frac{BB}{JB} = PA - PB$$

In conclusion, the ability to discriminate is important in approach to score who gets more correct answers and it directs to more of the ability in question, and to score who gets fewer correct answers which is related to less of the ability in question.

**Item Distractors**

In particular, distractors have been designed primarily to draw uninformed or unskilled learners away from the correct answer (Popham, 2000). It reveals the purpose of the distractors is to appear as plausible solutions to the problem for those learners who have not achieved the objective being measured by the test item. Conversely, the distractors must appear as implausible solutions for those learners who have achieved the objective. Only the answer should appear plausible to these learners.

A high-quality item should have distractors. If a distractor is not selected with the frequency of the others in the same item, it detracts from the overall quality of the item and must be either eliminated or revised (Airasian, 2001).

Arifin (2009) stated in getting information of the effectiveness and appropriateness can be gained through these formulas and interpretations:

$$P = \frac{E}{Je} \times 100\%$$

$$ID = \frac{Upper\ group - lower\ group}{Total\ number\ of\ learners}$$

There are Excellent, Good, Deficient, Poor, and Very Poor as interpretation in deciding the effectiveness result. In appropriateness of distractor, if the result is positive value, means negative result because the upper group who masters the test chooses more the distractors than lower group. Indeed distractors are settled to cheat the learners who don’t master the test.

**RESEARCH DESIGN**

This research was designed into descriptive quantitative design. The subject in this study was the ninth graders of SMPN 7 Pasuruan in academic year 2016/2017. There were 8 classes from A class to H class. Each class consisted around of 32 learners with 253 total participants. Then, the object of the study was the first English simulation test with all codes; code 1, code 2 and code 3 with 50 items in every code. The researcher chose this test to know its quality which made by the teacher of the ninth grade in SMPN 7 Pasuruan. There were 2 kinds of instrument which used, those were, documentation consisted of three codes of first English simulation test, 253 learners’ answer sheets, key answers of every code, and lattice of arranging this simulation test. The researcher got them by asking the headmaster of SMPN 7 Pasuruan and the researcher was allowed to copy it. This test was held on February 17th, 2017. And the researcher had demanded these documents in the same date before the test would be corrected at the same day of administering the simulation test. So the researcher corrected all learners’ answer sheets by himself manually. Next instrument was human instrument. In this research, human in-
struments were the researcher himself because the researcher had the experience to analyze data to know characteristics of test as what he has ever done in Second Language Acquisition (SLA) lecture. In this research, the researcher applied data codification to keep personality of subjects as the ethical code in this research and to make easier in analyzing data. The code was arranged based on learners’ initial name, and their numbers of test. For instance, MYM/07-H; MYM was the initial name of Muhammad Yusuf Maulana, and 07-H was his number of test.

The researcher generally decided to use Ms. Excel in presenting several steps to measure validity, reliability, item difficulty, discriminating power and item distractor of this simulation test:

**Step to decide Validity**

The researcher tried to analyze content validity of each code by matching the items with lattice. The researcher did these following steps: (1) Making a table to show what kinds of material and put basic competence. (2) Placing each items of each code in appropriate column. (3) Counting the sum and the percentage of items in every material. (4) Concluding the result into category.

**Step to Determine Reliability**

The researcher used Kuder-Richardson 20 formula (K-R20) in estimating inter-item reliability which is known as internal consistency. The researcher used K-R20 in these following steps: (1) Making tabulation of learners’ score in each item of each code using dichotomous scoring. (2) Measuring central tendency, and variability of learners’ score of each code as additional information. (3) Measuring reliability by using K-R20. (4) Checking the result of reliability with category and determining the skew.

**Step to Determine Item Difficulty**

The difficulty index represents the easy or difficult degree of every item in test. The researcher here concerned on using that Dubios’ formula in assemble steps: (1) Arranging tabulation of learners’ score to show learners’ response about each item using dichotomous scoring. (2) Analyzing item difficulty. (3) Interpreting the result into its quality and making a chart to show percentage of each quality. (4) Deciding the items need to be revised (0.1 > P > 0.9).

**Step to Determine Discriminating Power**

The researcher analyzed data in this study using Djiwandono’s formula and five criterions in these steps: (1) Separating all learners into two classes; Upper class and Lower class. (2) Measuring discriminating power with formula. (3) Classifying the result of discriminating power into five criterions and presenting the result through the chart of percentage.

**Step to Analyze Item Distractors**

Basically, there are only four steps in analyzing the work of item distractors of its effectiveness and appropriateness, included: (1) Calculating the item analysis statistic. (2) Computing the proportion of the learners in High and Low who selected each option (A,B,C,D). (3) Measuring and interpreting the effectiveness of item distractors. (4) Measuring and classifying the appropriateness of item distractors.

**FINDING AND DISCUSSION**

**Finding of the Research**

As the additional information in analyzing test, the researcher presented the entire result of learners’ accomplishment to this test. The standard score of accomplishing this simulation test was 55. It could be perceived that there are 258 learners of ninth graders but 5 learners did not join the test. There was 52,33% of learners were not able to pass the test. Only 45,74% of total learners who passed.

**Validity of Simulation Test**

(1) Making a table based on materials in lattice of simulation test.
Based on lattice, there were 3 kinds materials with their branches such as transactional/interpersonal text; apologizing, hope, pray, intention, habit, requirement, and congratulation, short-functional text; announcement, notice, label, invitation, advertisement, and text types; descriptive, recount, narrative, procedure, and report. Apologizing, and habit material included of seventh grade syllabus. Announcement, invitation, and descriptive material included of eighth grade syllabus. The others included of ninth grade syllabus.

(2) Placing the item of each code in appropriate column
The researcher found there were item number 36 and 38 which could not be placed in any materials of lattice. Then, the researcher found also there were material of pray and habit material which had no item in any code of this simulation test.

(3) Counting the sum and the percentage of items in every material in lattice
Concluding the result into category
Code 1 had 96%, code 2 had 100% and code 3 had 100%, it can be concluded that all of codes in simulation test had good category of content validity.

Reliability of Simulation Test
There were 0.67 and 0.64 indicated reliability of code 1 and 2. Both coefficient was in high category of reliability. Meanwhile, code 3 showed 0.37 of reliability which had a meaning of low category. In making the diagrams, code 1 and code 3 didn’t direct to left or right but to the middle means “normal skew”. In the other hand, code 2 showed that learners’ score almost directed to left side of the diagram. This means code 2 had “positive skew”.

Item Difficulty of Simulation Test
(1) Arranging tabulation of learners’ score. This tabulation gave the information about the relationship between learners’ correct answers toward deciding certain level of difficulty of each item.
(2) Analyzing item difficulty
(3) Interpreting the result into its quality and making a chart to show percentage of each quality. Code 1 had 16 items (32%) as easy item, 26 items (52%) as moderate item with, and 8 items (16%) as difficult item. Code 2 had 17 easy items (34%), 22 moderate items (44%), and 11 difficult items (22%). Then, Code 3 had 11 easy items, 24 moderate items and 15 difficult items with each percentage of 22%, 48% and 30%.
(4) Deciding the items need to be revised
The researcher found out the items that must be taken place and revised in each code in this simulation test. Code 1 had 12 too easy items and 2 too difficult items, code 2 had 6 too easy items and 1 too difficult item, Code 3 had 4 too easy items and 5 too difficult items. In the meantime, the researcher concluded the materials which had more the role in administering too easy or too difficult items in every code. The first material was descriptive text. For instances, item 8 and 25 of code 1, item 9 and 30 of code 2, item 23 of code 3. Then, next material was report such as item 19 of code 1, item 48 and 50 of code 2, item 25 of code 3. Both material could be found in every code in this simulation test.

Discriminating Power of Simulation Test
(1) Separating the learners into two classes; Upper group and Lower group
The samples were taken 30% as the upper group and as the lower group. Code 1 had 26 learners. Code 2 had 25 learners. Then, Code 3 had 24 learners as a sample.
(2) Measuring discriminating power

<table>
<thead>
<tr>
<th>No.</th>
<th>Material</th>
<th>Code 1</th>
<th></th>
<th>Code 2</th>
<th></th>
<th>Code 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Σ</td>
<td>%</td>
<td>Σ</td>
<td>%</td>
<td>Σ</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Apologizing</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Hope</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Pray</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Intention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Habit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Requirement</td>
<td>6</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Congratulation</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Announcement</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Notice</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Label</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>Invitation</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>Advertisement</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Descriptive</td>
<td>6</td>
<td>12</td>
<td>9</td>
<td>18</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>Recount</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>Narrative</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>16</td>
<td>Procedure</td>
<td>7</td>
<td>14</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Report</td>
<td>8</td>
<td>16</td>
<td>12</td>
<td>24</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>48</td>
<td>96</td>
<td>50</td>
<td>100</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1 The sum and percentage of items in every material
(3) Classifying the result and presenting through the chart.

The data which was shown from code 1 was 11 good items (22%), 13 satisfactory items (26%), 12 poor items (24%), 6 items of no discriminating power (12%), and 8 negative items (16%). Code 2 had 5 good items (10%), 23 satisfactory (46%), 11 poor items (22%), 3 items of no discriminating power (6%), and 8 negative items (16%). Then, code 3 indicated that it had 3 good items (6%), 19 satisfactory items (38%), 15 poor items (30%), 2 items of no discriminating power (4%), and 11 negative items (22%).

Talking about poor, no discriminating power and negative criterions, the researcher analyzed the relationship among them with too easy or too difficult items in item difficulty. For instance, in code 1 there were item number 4, 6, 8, 15, 16, 19, 25, 28, 31, 34, 35, 36, 46, 49 as too easy or too difficult items, they all were matching with poor, no discriminating, or negative criterion in discriminating power. It happened also in code 2 with the item number 9, 13, 30, 35, 41, 48, 50. Then code 3 which presented too easy and too difficult items of 4, 11, 14, 17, 23, 25, 32, 44, 31.

Item Distactors of Simulation Test

(1) Calculating the item analysis statistic
The researcher put all the learners’ responses to each distractor (A, B, C, D) in each item.

(2) Computing the proportion of the learners in High and Low
This was done by arranging the score from the highest to the lowest. In code 1, the researcher used 26 learners, code 2 had 25 leaners, and 24 learners were the number of learners to each upper and lower group.

(3) Measuring and interpreting the effectiveness of item distractors
The researcher focused on analyzing the item distractors were classified in poor or very poor, because they were not as workable distractors and too dominated in use. Obviously, they were needed to be revised. The researcher revealed the data of code 1 that there were 78 poor distractors (52%), and 29 very poor distractors (19%) in the upper group. There were 60 poor distractors (40%) and 19 very poor distractors (13%) in the lower group. Code 2 had 81 poor distractors (54%) and 27 very poor distractors (18%) in the upper group. It had also 46 poor distractors (31%) and 21 very poor distractors (14%) in the lower group. Thus, in code 3, the researcher found 51 poor distractors (34%) and 17 very poor distractors (11%) in the upper group, and there were 32 poor distractors (21%) and 13 very poor distractors (9%) in the lower group.

(4) Measuring and classifying the appropriateness of item distractors
The researcher here wanted to focused on measuring the unappropriated distractors that had positive value (+), because they were needed to be revised. The table below was used to presented what the researcher wanted.

In deciding the appropriateness from 150 of total distractors in each code, table 4.12 stated that 32 distractors of code 1 (21%) didn’t work at desired place. They didn’t work more in appropriate group but worked more in wrong group. There were 25 distractors of code 2 (17%) which worked more in upper group than in the lower group. Code 3 had 39 distactors (26%) which served bad appropriateness in working.

DISCUSSION

The researcher came to next steps to determine the test quality by discussing the result with the theories from the experts.

Validity

In finding, code 1 showed 96% of content validity. Code 2 and code 3 had the same percentage, that was, 100% of content validity. According to Roa (2009), the test has good content validity if it puts all materials in curriculum which are intended to measure. This theory strengthens the researcher’s conclusion about the content validity of code 1, code 2 and code 3 were good. It fulfilled one of characteristics as a good test. Although, theory from Lukke (2014) revealed test can be said to have real high validity if the proportional of spreading materials accentually highlighted. The researcher found some materials of lattice werent put into the test. This meant although this simulation had good content validity but the proportional of spreading materials wasn’t emphasize by the test maker.

Reliability

Code 1 had 0.67 coefficient as high reliability, code 2 had 0.64 coefficient as high reliability. Meanwhile, code 3 had 0.37 coefficient as low reliability. Fulcher & Davidson (2007) stated that whenever a test is administered, high reliability gives the assurance that the result could be replicated the same if the same individuals are tested again under the simi-
lar circumstance. In this simulation test, code 1 and code 2 fulfilled characteristic as a good test. In the other hand, reliability of code 3 could not be considered as completing characteristic as a good test.

In former study, imperfect reliability accentually needs to be revised because of the unknown factors (Indrawati, 2016). The researcher here, tried to analyze the factors which influence the reliability of each code especially of code 3 not to achieve very high category theoretically. According to Osman (2010), there are 5 factors in influencing the result of reliability. The factors which the research knew were scoring, standard and length that didn’t have much role to impact the reliability because all of codes had the same way to score, had the same standard and had the same numbers of item. So, the other factors were administration and instruction probably had the answer why the reliability of this simulation test especially code 3 wasn’t satisfying.

Item Difficulty

The researcher discussed the result of measuring item difficulty in finding to the items and got the result that this test had good item difficulty because all codes did not direct to easy or difficult level. The theory from Purwanto (2009) supports also this result, by stating a good test item should have moderate difficult level. So all codes of this simulation test had good characteristic of item difficulty. However, Arikunto (2008) agrees that the item outside of this proportion \(0.1 < P < 0.9\) must be changed because it is considered as too easy and too difficult item. In deeper analysis of item difficulty, the researcher found there were 2 materials which had a role in creating too easy and too difficult items to all codes in this simulation test, those were, descriptive and report. Based on the theory of item difficulty which is revealed by Katz et al (2002), there were 4 factors which influence item difficulty. The researcher connected what he found to this theory and got the result that descriptive and report materials might be too challenging for the learners’ ability to understand the questions.

Discriminating Power

According to the view of previous research, a good test is supposed to be at least in satisfactory category to show the decimating power of the test (Ali, 2015). The researcher concluded that based on the data in finding that all codes of this test had more satisfactory items than other criterions. So, the result of measuring discriminating power of all codes in this simulation test completed the characteristic to be a good test.

The researcher found all too easy or too difficult items were suitable with bad criterions of discriminating power in this simulation. This meant poor, no discriminating, and negative items were arranged by too easy and too difficult items. It proved the factor of discriminating power which revealed by Ketz et al (2002), discriminating power is greatly influenced by item difficulty.

Item Distractors

According to Airasian (2001), if a distractor is not selected with the frequency of the others in the same item, it distracts from overall quality of the item and must be either eliminated or revised. In the upper and lower group of code 1 didn’t showed the effective distractors. In code 2, the distractors in upper group worked ineffectively. But the lower group showed more than a half distractors showed the effectiveness. Code 3 had workable distractors in upper or lower group because the amount of ineffective distractors were fewer than the effective distractors.

The appropriateness of distractors was as the additional analysis that makes this research different to Ali’s study (2015) and Indrawati’s study (2016). According to Popham (2000), in particular, distractors are designed primarily to draw uninformed or unskilled learners away from the correct answers. This theory tells that distractors must be appropriate to the lower group than the upper group. All codes in this simulation test had more appropriate distractors than inappropriate distractors based on percentages above.

Having said these, the researcher concluded that code 1 didn’t have effective distractors but it had appropriate distractors. Code 2 couldn’t be considered to have effective distractors but it had appropriate distractors. Code 3 had both effective and appropriate distractors. Test can be perceived as a good test if it has both effective and appropriate distractors. Only code 3 which fulfilled the characteristic of item distractors as a good test.

CONCLUSION

The conclusion here is stated in accordance to the result of analyzing data. In analyzing validity, the researcher found code 1 of this simulation test had 96%. In code 2 and code 3 had 100%. All code in this simulation test are categorized in good validity. Although, it can not be confessed as real high validity because all codes did not emphasized good proportion of spreading the materials. In measuring
reliability, code 1 had 0.67 and code 2 had 0.64. They were categorized as high reliability. In the other hand, code 3 had low reliability with 0.37. These results are not satisfying result especially in code 3, they might be influenced by the factors of administration or instruction. In knowing the item difficulty, the researcher informed the data of all codes were dominated by moderate items. Code 1 had 26 (52%), code 2 had 22 (44%), and code 3 had 24 (48%). All codes did not direct to easy or difficult item. This means all codes in this simulation test have good test items. But, there were some items of each code which needed to be revised because they included in too easy or too difficult items. These items were mostly in descriptive and report material. This happened because of the factor of too challenging for the learners’ ability in understanding those materials. Discriminating power of all codes indicated more satisfactory items such in code 1 had 13 (26%), in code 2 had 23 (46%), and in code 3 had 19 (38%). All codes of this simulation test have good items to differ the learners’ ability who mastered and who did not. The researcher found also that all too easy or too difficult items in item difficulty influenced poor, no discriminating power, and negative items in this discriminating power. It can be understood that discriminating power is influenced by the item difficulty. The effectiveness and appropriateness are emphasized in analyzing item distractors. Code 1 had 71% in the upper group and 57% in the lower group, code 2 had 72% in the upper group and 45% in the lower group, code 3 had 45% in the upper and 30% in the lower group of uneffective distractors. In understanding the effectiveness of item distractors, code 1 and code 2 had more unworkable distractors, but code 3 had more effective distractors. Knowing inappropriate distractors by revealing shows code 1 had 32 (21%), code 2 had 25 (17%), and code 3 had 39 (26%). In analyzing the appropriateness of item distractors, it is concluded that all codes of this simulation test had good distractors which worked appropriately. Having said these, the conclusion of first English simulation test for ninth graders in SMPN 7 Pasuruan in academic year 2016/2017 are taken to each code. All codes have only one lack in fulfilling the characteristics as a good test. Code 1 and 2 lack in the effectiveness of items distractors, code 3 lacks because of having low reliability.

Having finished doing this research, the researcher realizes the lack of this study especially in using instruments to prove the factors which influenced the quality of test. the researcher gives also some suggestions to the test-makers should understand the importance of making a good test in measuring learners’ achievement. Putting all characteristics in arranging a good test is not easy to do, but by giving attention to avoid factors that can influence characteristics can make a good test easier to be created. It will better if the test is simulated first to check whether it fulfills all characteristics or not before it is administered. To other researchers who have willingness to conduct the same design as this research, it is suggested that not only to concern on analyze or measure the characteristics of a good test, but also focus on knowing the factors which influence to fulfill the characteristics. This research can be a guidance to conduct it. However, the other researchers’ studies hopefully are capable of using more data instruments to support and expand the finding.

REFERENCES


