AN ANALYSIS OF STUDENTS READING FINAL EXAMINATION BY USING ITEM ANALYSIS PROGRAM ON ELEVENTH GRADE OF SMA NEGERI 8 MEDAN

Dewi Manalu
Kammer Tuahman Sipayung
Febrika Dwi Lestari
English Department, University of HKBP Nomnensen,
Sutomo St. No. 4A, Medan, Indonesia
Email: dewimanalu54@gmail.com

ABSTRACT

The purpose of this study was to determine the quality of the reading final examination in SMA N8 Medan grade eleventh in terms of reliability, level difficulty, discrimination power and level of distractor. This research is qualitative-quantitative research. The subject of research are the grade XI of SMA N8 Medan. Data is analyzed by ANATES program version 4.0.9. The analysis shows that: (1) items of multiple choice question that can be said as valid are 14 items (56%), while the invalid items amounted to 11 items (44%). (2) Items multiple choice question can said as reliable because it equal 0,90 so it can said reliable. (3) items of multiple choice questions that categorized easy are 3 items (12%), satisfactory category 7 (28%) difficult category 2 (8%) and the other calculate categorized very easy 3 items (12%) and very difficult 3 (12%). (4) items of multiple choice questions that categorized poor are 12 items (48%), average category are 2 items (8%), good items category 1 items (4%) and excellent items are 8 items (32%).

Keywords: Test Item Analysis, Reliability, Level of Difficulty, Discrimination power, Level of Distractor.

Introduction

Education is needed by everyone, it can be said that education is experienced by all people. Education plays an important role in ensuring the survival of the nation and the state. Education as a means to improve the Human Resources (HR). Human Resources quality can bring the progress of Science and Technology in all aspects of life and bring people to a more advanced civilization and logical mindset (Rukmana, 2013). Human resource development can be enhanced by improving the quality of education in schools, which is determined by various factors. These factors are teachers, students, environment, infrastructure, learning time and learning process. Human Resources quality will appear along with a good quality of education as well. Until recently, the education component that is used, needs to be improved especially in the evaluation system of learning outcomes. Evaluation of education according to UU No. 20 Year 2005 are the activities of controlling, underwriting and, determining the educational quality of the various components of education at every track, level and type of education as a form of educational responsibility. One of the components that have to be evaluated is the learning outcomes of students. According to Sridadi (2007) assessment of student’s learning outcomes intended as an attempt to collect a variety of continuous and comprehensive information about the process and the outcomes of learning that have been achieved by the students through the teaching and learning activities as the basis for determining the next steps. Based on Government Regulation (PP) No. 19, 2005 on National Education Standards, particularly Article 63 paragraph 1 states that education in primary and secondary level consists of assessment of learning outcomes by educators, educational units and government. Article 64 Paragraph 1 states that the assessment of learning outcomes as intended in Article 63 paragraph 1, point (a) is done on an ongoing basis to monitor the process, progress, and improvement of the results in the form of daily tests, midterm examination, the examination in the end of the semester, and examination in order to go to the next grade in terms of good or poor, healthy or sick, clever or stupid, high or low, and so on (Djaali and Pudji Mulyono, 2008). One technique that is commonly used for the assessment is by using the test. The test is a procedure that can be used to determine or measure something in accordance to the way and the rules that have been set (Suharsimi, 2013: 3). Evaluation is a process of collecting data to determine the extent in terms of what and how the educational goals have been achieved (Suharsimi, 2013: 3). Evaluation can be used as a benchmark in decisions making about the object that will be evaluated. Measurement is an attempt to compare a particular casewith so its quantitative (Arikunto and Jabar, 2004). While the assessment is making a decision on something
According to Suharsimi (2013: 73), the data can be said to be valid if it is in accordance with the actual situation. If the data is valid, the instrument used is valid.

Language is a set of rule. Realization of rule can be recognized through grammar. Sipayung (2018) stated that Teaching English is a complex enterprise. He is doing it in classroom. In a simple synoptic, a teacher is greeting students, explaining the matter, ask learners participants and do the homework. This has been teaching trade and a tradition. But, when a search of teaching is done, it is an interction between a teacher and his community, the classroom subject, the pupil. Teacher made test as a level of students understood. One aspect in learning process is do an evaluation to the students. It is done to make sure whether the learning process has been running well over the term. In education field, evaluation has an important role, because it shows the results of learning program. The objective of the evaluation itself is to help the teacher ascertain the degree to which educational objectives have been achieved, to review the effectiveness of teaching method and to help the teacher know his pupils as individuals.

According to Ahman and Glock evaluation is the systematic process of determining the effectiveness of educational endeavors in the light of evidence. Once assessment information is collect, teacher use it to make a decision, reflection or judgment about pupils, instruction or classroom climate. The evaluation is defined as the process in making decision and solution for education process based on the result of tests, other assessments, and others reports Brown & Priyanvada (2010:9). For English language education, the evaluation is conducted in many aspects of education such as curriculum, strategies of teaching, references and test items. The good evaluation is believed will have good benefits in national education because by evaluating the system of education, the government can improve the quality of education system.

Developing a test is a complex and reiterative process which subject to revision even if the items were developed by skilful item writers. Many commercial test publishers need to conduct test analysis, rather than trusting the item writers judgement and skills to improve the quality of items that need to be proven statistically after trying out was performed. This tudy is a part of test development process which aim to analyse the reading test items.

Testing refers to an effort to measure the result of students learning in teaching learning process. Consequently, the teachers should have an ability to arrange and good test and analyze of a test. Therefore, the accuracy and the carefulness of teachers may have a big impact on the increase the quality of teaching particularly in giving the judgement of students ability. This information is ver useful for both students in their learning and the teacher in their learning. It can be feedback for the teacher, who have responsibility to meet the instructional objectives, while for getting the data for the sake of evaluation purpose, one them is by using a test. The test items are supposed to be well constructed so it can be used efficiently. To be an effective test, it has to fulfill the criteria of a good test, they are validity, reliability and practicality. The test is recognized as valid if can measures what supposed to be measure, it illustrates their performance. Related to the importance of the evaluation, it is necessary to consider that the test should be well constructed. As a means of evaluation, a test is administrated to get information about the students improvement and to measure the result of the teaching
learning process. Test a test activity which is held at the end of teaching learning process in one semester. That is why, that test is a kind of test which is intended as a feedback from the students and also as a result of teaching from the teachers in one semester. This information will be used to consider and to decide several rules not only for the student’s but also for the teachers in increasing the quality of teaching learning process. And the English test is made by MGMP (Musyawarah Guru Mata Pelajaran). While MGMP itself consists of a team who has responsibility to design a test for each subject, it means that the semester test items are rarely analyzed by the teachers after they are tested.

To analyze the semester test items, there are some criteria of a good test according to some expert. A good test should have (1) Validity, (2) Reliability, (3) Level of difficulty, (4) Discrimination Power, and (5) The Quality of Options. This research was concerned with the whole test items designed by MGMP. This includes test analysis and item analysis. Test analysis is administered to determine and describe such criteria as face validity, content validity, construct validity, and reliability. And the item analysis is used to determine about the level of difficulty, discrimination power, and the quality of options. Shohamy (1985:3) supports that a test is a sample of knowledge and needs to be a good representation of it. It means that, what should be tested just a sample of behavior or knowledge, not the whole or behavior what the teachers have taught the students have learned because it is also impossible to measure all of the students’ abilities. The things that should be taken into account is the sample must be representative in the sense which is tested, it should reflect the knowledge that has been taught. The test that has been analyzed was an achievement test and it was designed by MGMP. Achievement test tried to investigate the students’ achievement based on the objective of a given material. Achievement test (Harrison as quoted by Hayatunnisa, 2003:8) tries to evaluate the test takers’ language in relation to a given curriculum or material which the test-taker had gone through in a given course. It is intended to show the standard which the students have reached in relation to other students at the same stage.

A good test should fulfill certain criteria. There are four criteria of a good test according to some expert; they are validity, reliability, level of difficulty, and discrimination power. Validity refers to the extent to which an instrument really measures the objective to be measured and suitable with the criteria based on Hatch and Farhady (1982:250). Another words, a test can be said to be valid to the extent that it measures what it is supposed to measure. If the test is not valid for the purpose for which its design, the scores do not mean what they are supposed to mean. Reliability refers to the consistency of measurement that is, to see how consistent test scores or other evaluation results are from one measurement to another (Grondlund, 2000:193). It means that a test is administered to the same condition on different occasion, the extent that it produces different result, it is not reliable. Discrimination power is an aspect of item analysis, discrimination power tells about which is the item that discriminates between the upper group students and the lower group students. Shohamy (1985:81) states that discrimination index tells about the extent to which the item differentiates between high and low students on that test.

Difficulty level is one of kind of item analysis. Level of difficulty was concerned with how difficult or easy the item for the students. Shohamy (1985:79) states that difficulty level relates to how easy or difficult the item is from the point of view of the students who took the test. It is important since test items which are too easy can tell us nothing about differences within the test population. If the item too easy, it means that most or all of the students obtained the correct answer. In contrast, if the item is difficult, it means that most or all of the students get it wrong. The quality of options is a distribution of test in decided alternatives on a multiple-choice test. It is obtained by calculating the number of tests who choose the alternatives A, B, C, or D or those who do not choose any alternatives. From this way, the teachers would be able to identify whether distracters function well or bad.

A good evaluation (test) is important instrument to know whether the students need a help or not. Test have been widely use to demonstrate level of proficiency of the students, and at the same time function as policy instruments to implement educational standards stated by Phakiti & Roever (2011:29). This study is a part of test development process which aims to analyse the reading comprehension test items. A teacher who is competent in terms of pedagogics must be able to make quality questions. The quality of the questions is viewed from the aspects of difficulty items, item discrimination power and level of distractors.

Item analysis is a crucial part in a test development process as it functions to provide information about items that should be improved in terms of quality for later tests or even be eliminated due to misleading. According to Grondlund (1977) there are some benefit to do analysis of the test items. First, it is provide useful information for class discussion of the test, second, it provides data that helps the students improve their learning. Third, it provides insight and skills that lead to the preparation of better tests in the future. Item analysis is process which examines students responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. It will show whether the test appropriate or not for the students. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors skill in instruction, and identifying specific areas of course which need greater emphasis clarity.

Reading Comprehension
Reading comprehension can be defined as the ability to understand vocabulary in order to paraphrase and make a summary of information from the text based on Manarin et al (2015). It is the activity to reconstruct a message from written symbols to a form of language, and it involves many cognitive processes and combines both decoding process and inferential activity and so that the readers can really comprehend the text. In the context of English as a foreign language, reading English textbooks becomes a big issue, particularly for students with non-English background. Reading skills seems to be the big problem to the students because most of them find reading English text is difficult. reading comprehension is part of critical reading that can be a determinant to academic success. Lowes et al (2004) argues that reading is essential to understand basic concept of a subject to gather information for completing assignments, and to improve English skill, particularly to increase vocabulary. One of many characteristics of reading at higher education is critical reading which involves such features as identifying patterns of textual elements, distinguishing main and supporting ideas, making credible evaluation and arguments and making relevant inference about the text.

Reading Comprehension Test

According to Brown test is a tool that serves to make decisions related curriculum and other areas. Brown (2004) points out that a test is a way of measuring one’s ability, knowledge or performance in a given domain. In language testing context, most test measure test takers competence, such ability to perform language skill to speak, write, listen or read to one subset of language. These performance based test sample the test takers actual use of language which infers general competence. A test of reading comprehension, for example, may consist of several short reading passages each followed by a limited number of comprehension questions. From the result of the test the examiner may infer a certain level of general reading ability. Constructing a test is not a simple task. It involves a science and art of many complex task, such as planning, test preparation an administration, scoring, statistical analysis, and test result report (brown, 2004; downing, 2010). One of crucial stages in test development is statistical analysis of a test.

Reading Assessment

Assessment is a systematic approach to collecting information and making inferences about the ability of a student or the quality or success of a teaching course on the basis of various source of evidences. To collect the information of the students improvement in reading comprehension, some assessment of reading are essential to be carried out. Unlike speaking and writing, the reading process and product cannot be seen and observed specifically. For this reason, all assessment of reading must be carried out by inference. Some consideration are needed in designing assessment of reading, such as the types of genres of written text, the components of reading ability and specific tasks.

Furthermore, the types of reading performance will influence the assessment tasks as well. Brown lists a number of possible tasks for assessing perceptive, selective, interactive and extensive reading. Related to the spesifications of reading in this study that is assessing students interactive reading performance so there are three types of possible assessment tasks applied as follow 1. Multiple-Choice Multiple choice in this study provides not only the vocabulary and grammatical items but also the context to assess the students understanding of information in the text. The context is presented by putting a pair or part of a text followed by questionsin Method

This study aimed particularly process of analysing multiple choice questions in reading skills, to improve item quality. Twenty-five of multiple choice question had been tested to get the evidence on item quality. The items were one correct answer type, having a stem and four options, one of them being correct and the other three being “distractor”. Based on the students responses, the test items were then analysed using ANATES Software.

Analysis data refers to find and set which the students have to respond correctly.2. Impromptu Reading Plus Comprehension This type of assessment involves impromptu reading and responding to questions. It is commonly used in proficiency test. In this test, students are provided a reading passage followed by questions and have to responds to the items. The set of questions in impromptu reading covers the comprehension of some features of reading: (1) main idea (2) expressions idioms/phrases in context (3) interference (4) grammatical vocabulary in context. These spesifications and the questions are in line with strategies of effective reading:skimming for main idea, scanning for details, guessing word from context, inferencing using discourse makers etc which are assessed in this study. 3. Short Answer Task In this type of assessment, reading passage is presented, and the students read questions that must be answeredin a sentences or two. The question might cover the same specifications similar to the impromptu reading. Those three types of assessments task have the combination of form-focused and meaning-focused objective. They cover the objective of reading assessment especially in the comprehension issues and embody the evidences of students reading comprehension.

Anates is a software for statistic calculation. Anates V4 develop by Drs. Kanoto M.Pd and the programer is Yudi Wibisono ST. This software very effective to calculate about reliability, item difficulty, item discrimination power, and level of distractor. The advantages in using this program is this program can be use to analyze the test item of multiple choice and it really helps for investigate the data.
collecting data process to be easily to understand in this study. The analysis quantitatively is doing to order investigate the test items from students answer or response and the key answers. Anates is a software for statistic calculation. Anates V4 develop by Drs.Kanoto M.Pd and the programer is Yudi Wibisono ST. This software very effective to calculate about reliability, item difficulty, item discrimination power, and level of distractor. The advantages in using this program is this program can be use to analyze the test item of multiple choice and it really helps for investigate the data.

**Findings and Discussion**

Based on the analysis, the reliability was 0.90, the value was categorized as a high it has a certain criteria in test reliability.

**Item Difficulty**

In testing item difficulty, the formula were applied:

\[ P = \frac{n}{JS} \]

\[ P = \text{difficulty index} \]

\[ B = \text{the number of students who got the item right} \]

\[ JS = \text{the total number of students who tried to answer the item} \]

How to use anates software:

1. Open the application and the application will ask how many question and how many students then fill it.
2. Input the data / score students one by one then save
3. Click on “process data” on the option
4. Then data will process and give the result based on item difficulty, item discrimination and level of distractor that you choose before.

**Table 1. Distribution of multiple choice question of final examination of SMA Negeri 8 Medan based on level difficulty**

<table>
<thead>
<tr>
<th>No</th>
<th>Difficulty Level</th>
<th>Item Number</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.70 - 1.00 (easy)</td>
<td>4,16,21</td>
<td>3</td>
<td>12%</td>
</tr>
<tr>
<td>2</td>
<td>0.31 – 0,71 (satisfactory)</td>
<td>1,2,5,8,12,17,20,24,25</td>
<td>9</td>
<td>36%</td>
</tr>
<tr>
<td>3</td>
<td>0.00 – 0.30 (difficult)</td>
<td>9,11</td>
<td>2</td>
<td>8%</td>
</tr>
<tr>
<td>4</td>
<td>0.00 – 0.00 (very difficult)</td>
<td>13,14,15</td>
<td>3</td>
<td>12%</td>
</tr>
<tr>
<td>5</td>
<td>1.00 – up (very easy)</td>
<td>3,6,7,10,18,19,22,23</td>
<td>8</td>
<td>32%</td>
</tr>
</tbody>
</table>

**Item Discrimination**

Item discrimination has a significant role to examine if an item is low or high quality. In testing item discrimination, the formula were applied:

\[ D = \frac{BA - BB}{JA - JB} \]

\[ D = \text{Discrimination Power} \]

\[ BA = \text{the number of students in upper group who answer correctly} \]

\[ BB = \text{the number of students in lower group who answer correctly} \]

\[ PA = \text{index of difficulty in upper group} \]

\[ PB = \text{index of difficulty in lower group} \]

Based on the analysis result, the number of multiple choice questions which have poor categories are 12 items (48%), average category are 2 items (8%), good items category are 1 items (4%) and excellent items are 8 items (32%).

**Table 2. Distribution of multiple choice question of final examination of SMA Negeri 8 Medan based on discrimination power**

<table>
<thead>
<tr>
<th>No</th>
<th>Difficulty Level</th>
<th>Item Number</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00 – 0.19 (Poor)</td>
<td>3,5,6,7,10,13,14,15,16,18,19,23</td>
<td>12</td>
<td>48%</td>
</tr>
<tr>
<td>2</td>
<td>0.20 – 0.39 (average)</td>
<td>4,22</td>
<td>2</td>
<td>8%</td>
</tr>
<tr>
<td>3</td>
<td>0.40 – 0.69 (good)</td>
<td>9</td>
<td>1</td>
<td>4%</td>
</tr>
<tr>
<td>4</td>
<td>0.70 – 1.00 (excellent)</td>
<td>1,2,8,11,12,17,20,25</td>
<td>8</td>
<td>32%</td>
</tr>
</tbody>
</table>

**Distractor Analysis**

The distractors can function well if at least chosen by 5% of all learners who participate in test. Criteria for assessing the use of distractors can adapt based on Arikunto (2005 :220) are:

a. Acceptable, because it has been chosen at least 5% the total number of students who try to answer. A good distractor can distract more
students from lower group than students from upper group.

b. Refused, because it does not exist by the students to choose the distractor.

c. Revision, because it attracts more students from upper group than students in lower group. It is revision distractor if choose by less of 5% of the total number of students who try to answer. A distractor can be called revision only on the sentence structure. It can be re-write the effective revision.

For example:

Once, there was a family who had a baby. They also had a dog named Blackie and it was very smart and faithful. Blackie used to take care the baby while the family working in the Farm.

One afternoon, while they were working, they heard Blackie barking. Blackie was running toward them with mouth covered in blood. The husband shocked. He thought that Blackie killed the bab. Suddenly he took his grass knife and hit the dog until it died. They got home around, he found a die big Snake covered in blood. It seemed that blackie killed the Snake and he killed blackie.

1. Who was Blackie? Blackie was…
   a. The dog  
   b. The husband  
   c. The wife


<table>
<thead>
<tr>
<th>Option</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Distractor B are functioning poorly since it attracts only one students from lower group. These distractor are able to distract more students in lower group. Distractor C and D is unless attracts because no students in both of this group choose the option. This the analysis of distractor:

a. Distractor B = revise
b. Distractor C = refused
c. Distractor D = refused

test had high reliability with coefficient 0.90. based on the finding, many items were categorized as marginal and poor category in terms of the level of difficulty, discrimination power and level of distractor.

References


