## CHAPTER II

## REVIEW OF RELATED LITERATURE

### 2.1 The Definition of Evaluation, Measurement, Testing and Assessment

We are sometimes confused with the terms evaluation, measurement and testing because they are often used synonymously. To distinguish those meanings. Evaluation refers to the act or process of determining the value of something. ${ }^{8}$ In addition, Gronlund state that evaluation is qualitative descriptions of pupil behavior. Measurement as the process of quantifying the characteristics of person according to explicit procedures and rules. ${ }^{9}$ In line with Bachman, as Gronlund that measurement is a quantitative description of pupil behavior. Measurement means that the act or process of ascertaining the extent or quantity of

[^0]something. ${ }^{10}$ From those definitions, we know the differences between evaluation and measurement.

In educational process, measurement refers to the quantitative and evaluation refers to the qualitative. Nurkancana ans Sumartana also differentiate those terms. Measurement is used to answer the question "how much", whole evaluation is used to answer the question "what value".

Although evaluation and measurement are different, they are related to each other. Assessment of a program's outcomes or results (evaluation) is facilitated by measurement. ${ }^{11}$ In addition, Arikunto states that to evaluate something, we do measurement first. It means that when we are evaluating something, it should be based on measurement. For example, to evaluate student's reading ability, the teacher has to know the student's comprehension in reading. On the other hand, measurement will be useless if we do evaluate it. After we measure something, we do

[^1]evaluation on it. ${ }^{12}$ For example, if the students comprehend the reading text well, we can say that their reading ability are good.

For testing, test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual. ${ }^{13}$ A test can be considered to be a device typically used to find out something about a person. ${ }^{14}$ In addition, Arikunto that test is a device or a procedure which is used to find out or to measure something. Here, a test is used to measure the changing of individual's behavior as the goal of instruction. By giving a test the changing of their student's behavior. The objectives of language testing. ${ }^{15}$

1. To determine readiness for instructional programs.
2. To classify or place individuals in appropriate language classes.
3. To diagnose the individual's specific strengths and weaknesses.
4. To measure aptitude for learning.

12 Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 2
${ }^{13}$ Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. USA: Oxford University Press.pg. 20
${ }^{14}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes: Fundamental of Testing. USA: Harcourt Brace Javanovich.pg. 12
${ }^{15}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGraw-Hill.pg. 2
5. To measure the extent of student achievement of the instructional goals.
6. To evaluate the effectiveness of instruction.

Looking at the explanations above, we can conclude that a test is an instrument to give information about the student's ability and to decide something dealing either with the students or the teaching learning process.

We might be tempted to think of testing and assessing as synonymous terms, but they are not. Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated.

Assessment, is an ongoing process that encompasses a much wider domain. Whenever a student responds to a questions, offers, a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance. ${ }^{16}$

[^2]
### 2.2 Types of Test

### 2.2.1 Types of Test According to Its Role

According to its role in teaching, categories test into four categories. There are placement test, diagnostic, formative, and summative test. ${ }^{17}$

### 2.2.1.1 Placement Test

Placement test concerns with the student's entry behavior in a sequence of instruction. The goal of placement test is to determine the position in the instructional sequence and the mode of instruction that are more likely to provide optimum achievement for each student.

### 2.2.1.2 Diagnostic Test

Diagnostic test concerns with the student's persistent learning difficulties that are left unsolved by the standard corrective prescriptions of formative evaluation. In other word we can say that diagnostic test is a test of student learning difficulties during instruction. The primary aim of

[^3]diagnostic test is to determine the causes of learning problems and to formulate a plan for remedial action.

### 2.2.1.3 Formative Test

Formative test concerns with the student's learning progress during instruction and this test used to monitor learning progress. Its purpose is to provide continuous feedback to both students provides reinforcement of successful learning and identifies the specific learning errors that need correction. Feedback to teacher provides information for modifying instruction and for prescribing group and individual remedial work. Sine formative test is directed toward improving learning and instruction, the results are typically not used for assigning course grade.

### 2.2.1.4 Summative Test

Summative test concerns with the student's achievement at the end of instruction. It is designed to determine the extent to which the instructional objectives have been achieved and is used primarily for assigning course grades or for certifying student mastery of the intended learning costumes. The main goal of this test is not only for grading or
certifying student's mastery, but also for judging the appropriateness of the couse objectives and the effectiveness of the instruction.

In line with Gronlund, Johnson and Johnson state that summative test is conducted at the end of an instructional unit or semester to judge the final quality and quantity of student achievement and the success of the instructional program. ${ }^{18}$ In curriculum 2004, summative test is known as UAS (Ujian Akhir Semester) of final form test.

### 2.2.2 Types of Test According to The Test Maker

Beside types according to its ole in teaching, there are types of test according to the test maker. Categories test according to the test maker into two categories. ${ }^{19}$ They are standardized test and teacher-made test.

### 2.2.2.1 Standardized Test

Standardized test is a test which is made by professional testing services that the test is tried on first, analyzed, and revised before being used. Standardized test is designed to be used with thousands and

[^4]sometimes hundreds of thousands of subjects throughout the nation or the world, and prepared (and perhaps administered, scored, and interpreted) by team of testing specialist. ${ }^{20}$

In addition, standardized tests are prepared for nation wide use (usually commercial) to provide accurate and meaningful information on student's level of performance relative to others at their age or grade levels. ${ }^{21}$ They also state that such tests are usually constructed by subject matter specialists and experts on testing. To make the test scores comparable, the tests are administered and scored under carefully controlled conditions. It means that in standardized test, the test should be tried out, analyzed and revised before being used, UAN and SPMB are the examples of standardized test.

There are some characteristics of standardized test. According to Arikunto, the characteristics of standardized test are as follows: ${ }^{22}$

[^5]1. Based on the content and the general goal for the whole schools in the country.
2. In relation with general knowledge or capability.
3. Developed by professors, reviewer, and editors of test items.
4. Using items that are tried out, analyzed, and revised before being used for a test.
5. Having high reliability.
6. Having norms which represent the whole performance of schools in the country.

### 2.2.2.2 Teacher-Made Test

In teacher-made test, the test is made by the teacher himself of group of teachers without being tried on first, analyzed, and revised.Classroom test are generally prepared, administered, and scored by one teacher. ${ }^{23}$ In addition, Arikunto also states that the teacher-made testis

[^6]a test made by the teacher himself or group of teachers is using untried out, unanalyzed, and unrevised test items. ${ }^{24}$

The teacher-made test is used to measure his student's achievement on the objectives given after finishing the teaching learning progress. The teacher-made test is made by the teacher based on his or her own objectives and it is not tried out, analyzed and revised. ${ }^{25}$ Therefore, he also states that the teacher-made test has average or lower reliability than standardized test. UTS (Ujian Tengah Semester) or mid form test and UAS (Ujian Akhir Semester) or final form test are the examples of teacher-made test.

### 2.3 Forms of Test

There are some forms of test. According to Heaton there are two forms of test. They are subjective test and objective test are the terms are used refer to the scoring of the test. ${ }^{26}$ Objective tests usually only have one

[^7]correct answer, so they can be scored mechanically, while subjective test need scale for scoring the test.

### 2.3.1 Subjective Test

Subjective test or essay test requires students to express their own idea. In essay test candidates must think of what to say and then express their ideas as well as possible. ${ }^{27}$ In line with Heaton, Johnson and Johnson state that essay items require students to recall, select, organize, apply what they have learned and expressed it in their own words. ${ }^{28}$ It means that in subjective test or essay test the students are expected to think the answer and then express their ideas in a good arrangement. Essay items provide test takers with the opportunity to construct and compose their own responses within relatively broad limits. ${ }^{29}$

In the subjective test, the scorer's subjective judgment enters into the scoring. The scores differ from one scorer to another and from one time to another. It means that in scoring test, scorer's subjectivity influences the

[^8]test. Different scorer may produce different score. Subjective test are those that require an opinion, a judgment on the part of the examiner.

The opinion above lead to the conclusion of the strengths and the weakness of subjective test. Here are the strengths and the weaknesses of subjective test: ${ }^{30}$

The strength of subjective test are:
a. It is easy to construct the items.
b. It encourages the students to express their ideas and construct them in good sentences.
c. It is able to see how far the students master the material.

The weaknesses of subjective test are:
a. It has low validity and reliability because it is easy to know which knowledge has been mastered perfectly.
b. It lacks representative of all the materials that will be examined to the students.
c. It takes a long time in scoring.
d. It is difficult to score because it requires the scorer considerations.

[^9]
### 2.3.2 Objective Test

Objective test requires the students to choose the right answer ar give short answer. Objective tests are scored rather mechanically without need to evaluate complex performance on a scale. It means that in objective tests, the students are demanded to give short answer even only by choosing certain codes representatives of the answers available. ${ }^{31}$ Defines objective test as a short answer test.

In addition, Heaton states that objective test is referring to the scoring of the test that can be described as objective. In line with Heaton, Arikunto adds that the objective test is the test that can be scored objectively. ${ }^{32}$ It means that the student will get some score, no matter who examiners mark the test since it only has one correct answer.

The opinion above lead to the conclusion of the strengths and weaknesses of objective test. Here are the strengths and the weaknesses of objective test. ${ }^{33}$

The strengths of objective test are:

[^10]a. It can represent the materials that will be examined to the students.
b. It has high objectivity because it can avoid the sorer considerations.
c. It is easy to score and take a short time to score.

The weaknesses of objective test are:
a. It is much more difficult to construct than essay test items.
b. It tends to measure the cognitive aspect only.
c. It enables the students to speculate in choosing the correct answer.
d. It enables the students to cooperate to do the test.

In conclusion, because both subjective and objective test items have strengths and weakness, there is no best form of test. Therefore, the teacher should apply both of them in teaching learning process.

There are several types of objectives test. There are many varieties of there new types test, but four kinds are in most common use, true-false, multiple-choice, completion, matching. ${ }^{34}$ It will be discusses only the multiple-choice and the completion types.

[^11]
### 2.3.2.1 Multiple-choice Test

Multiple-choice test is a test where a testee has to select one correct answer from the option given. A multiple-choice item is usually set out in such away that the candidate is required to select the anwer from a number of given options, only one of which correct. ${ }^{35}$ In addition, Nurkancana and Sumartana state that a multiple-choice item is an item which consist of stem, which presents a problem situation, and several option, which provide possible solutions to the problem. ${ }^{36}$ The option include the correct answer and several wrong answer, called distracters, in which to distract those students who are uncertain of the answer. Briefly, it can be described as follows:

They usually. . . to work by train. stem
a. Gone
b. Went
c. Going
d. Goes

e. Go $\qquad$ correct option

[^12]In multiple-choice test, items should be constructed in such a way that students obtain the correct option by direct selection rather than the elimination of incorrect options. A good distracter will attract will attract more students from the lower group than the upper group. When item distracter attract more students from the upper group than the lower group, it is not a good distracter. And when item distracters do not attract both upper and lower group, it is a non function distracter.

The characteristics of a good multiple-choice test construction are as follows: ${ }^{37}$

1. Each multiple-choice item should have only one answer. This answer must be absolutely correct, unless the instruction specifies choosing the best option (as in vocabulary test)
2. Only one feature at a time should be tested. It has long been standard practice to test only one feature at a time, it is usually less confusing for the testee and it helps to reinforce a particular teaching point.
3. Each option should be grammatically correct when placed in the stem, except of course in the case of specific grammar test items.

[^13]4. All multiple-choice items should be at a level appropriate to the linguistic ability of testees. The contexts, itself, should be a lower level than the actual problem which the item is testing.
5. Multiple-choice items should be a brief and clear as possible (though it is often desirable to provide short contexts for grammar items).
6. In many test, items are generally arranged in rough order increasing difficulty. It is generally considered important to have one or two simple items to lead in the testee, especially if they are not familiar with the kind of test being administered.

### 2.3.2.2 Completion Test

Completion test is a test where the students have to fill in or complete a sentence or statement. In completion test, the students must construct their own response rather than choosing from among given choices. ${ }^{38}$ They fill in or complete a sentence from which a word or phrases has been omitted. Therefore, when we are dealing with completion test, we are filling in, giving a proper answer or completing a sentence or statement.

[^14]In constructing the completion test, the keys are as follows: ${ }^{39}$

1. To strike a balance between leaving out so much that the item becomes ambiguous and leaving out so little (or otherwise providing so many clues) that the items become so easy.
2. Avoiding instances where the grammar of the sentences helps determine the answer.
3. Completion items should have a single correct answer, preferably a word or short phrase.

### 2.3.2.3 True and False

Usually there are more true answers than false on most tests, if there is no guessing penalty, then guess. You have a $50 \%$ chance of getting the right answer. So the testee should read through each statement carefully, and pay attention to the qualifiers and keywords. If any part of the question is false, then the entire statement is false but just because part of a statement is true doesn't necessarily make the entire statement true.

Ideal test items:

[^15]- Critical content should be readily apparent and identified for analysis, avoiding cleverness, trickery, and verbal complexity
- Use simple, direct language in declarative sentences
- Present the correct part of the statement first, and vary the truth or falsity of the second part if the statement expresses a relationship (cause, effect--if, then)
- Statements must be absolute without qualification, subject to the true/false dichotomy without exceptions
- Every part of a true sentence must be "true"
- If any one part of the sentence is false, the whole sentence is false despite many other true statements.


## Limitations of using true-false items

True-false items:

- incorporate an extremely high guessing factor
- can often lead an instructor to write ambiguous statements due to the difficulty of writing statements which are unequivocally true or false
- do not discriminate between students of varying ability as well as other item types
- can often lead an instructor to favor testing of trivial knowledge


## True-False Test Items

A true-false item can be written in one of three forms: simple, complex, or compound. Answers can consist of only two choices (simple), more than two choices (complex), or two choices plus a conditional completion response (compound).

Sample true-false item:

- Simple

Conflict is essential in a play True False

- Complex
conflict is essential in a play True False Opinion
- Compound
conflict is essential in a play True False

If this statement is true, what makes it true?

### 2.3.2.4 Matching Test Items

In general, matching items consist of a column of stimuli presented on the left side of the exam page and a column of responses placed on the right side of the page. Students are required to match the response associated with a given stimulus.

Advantages in using matching items

- require short periods of reading and response time, allowing you to cover more content
- provide objective measurement of student knowledge
- provide highly reliable test scores
- provide scoring efficiency and accuracy

Limitations in using matching items

- have difficulty measuring learning objectives requiring more than simple recall of information
- are difficult to construct due to the problem of selecting a common set of stimuli and responses

Suggestions for writing matching test items

1. Include directions which clearly state the basis for matching the stimuli with the responses. Explain whether or not a response can be used more than once and indicate where to write the answer.
2. Use only homogeneous material in matching items.

Undesirable Directions: Match the following
1.

Impressionist
a. blue, red, yellow
2.

Pop Art
b. Claude Monet
3. primary colors
c. Andy Warhol
d. Claude Debussy

Desirable Directions: On the line to the left of each art style in Column I, write the letter of a representative artist from Column II. Use each name only once.

1. $\qquad$ Impressionist
a. Jackson Pollack
2. $\qquad$ Pop Artist
b. Claude Monet
3. $\qquad$ Abstract impressionist
c. Andy Warhol
d. Claude Debussy
4. Arrange the list of responses in some systematic order if possible (e.g. chronological, alphabetical)
5. Avoid grammatical or other clues to the correct response, e.g. avoid sentence completion due to grammatical clues.
6. Keep matching items brief, limiting the list of stimuli to under 10 .
7. Include more responses than stimuli to help prevent answering through the process of elimination.
8. When possible, reduce the amount of reading time by including only short phrases or single words in the response list.

### 2.4 Characteristics of a Good Test

Making a good test, a test maker should know the characteristics of a good test. So that the test is qualified enough to be given and can
represent the degree of the students' mastery over the language teaching materials have been thought.

All good tests include three qualities namely validity, reliability, and practicality. ${ }^{40}$ In this study, validity and reliability will be discussed because they are the most important characteristics of a good test.

A teacher, who wishes to use a good test to make an important decision about an individual or group, must be sure that the test possesses two absolutely essentials characteristics, validity and reliability. ${ }^{41}$

### 2.4.1 Validity

Validity refers to the extent to which the results of an evaluation procedure serve the particular uses for which they are intended. ${ }^{42}$ It means that validity of a test measures what it is supposed to measure. If the test is able to measure what its purposes, then the test has high validity. There are

[^16]three types of validity: content validity, criterion-related validity, and construct validity. ${ }^{43}$ However, only content validity will be discussed.

Content validity depends on a careful analysis of the language being tested and of particular course objectives. ${ }^{44}$ The test should be so constructed as to contain a representative sample of course, the relationship between the test items and the course objectives always being apparent. The test has content validity if the objectives stated in the curriculum. The sample of activities to be included in a test is as representative of the target domain as is possible. ${ }^{45}$ To know whether the test has content validity or not, the test should be compared with the materials states in curriculum. The test has high content validity if the test items cover the materials stated in the curriculum.

### 2.4.2 Reliability

Reliability refers to the consistency of measurement. It means that it shows the consistency of the test score or other evaluation results from

[^17]one measurement to another. The reliability of a test is a matter of how consistently it produces similar results or consistent reliability, then it can be said that the test has reliability. ${ }^{46}$ Published tests usually require test reliability of 0,85 or above while teacher- built tests are usually considered adequate with reliabilities of 0,60 or above. ${ }^{47}$

There are some factors effecting reliability of a test, those are: ${ }^{48}$
a. The extent of the sample of material selected for testing. It means that the test which has bigger items will be more reliable than the test which has small number of items.
b. The administration of the test. It means that the condition of administrating of the test will affect the reliability of the test.
c. The instruction. The clarity of the instruction will affect the students' comprehension to answer the test.
d. Personal factors, such as motivation and illness.
e. Scoring the test. It means that the objectives test is more reliable than the subjective test.

[^18]There are some methods to estimate reliability. Here, formula is used since it avoids troublesome correlations and it involves only the test mean and standard deviation, ${ }^{49}$ both of which are normally calculated anyhow as a matter of routine.

The formula is:

$$
\mathrm{r}=\left[\frac{\mathrm{N}}{\mathrm{~N}-1}\right]\left[1-\frac{\mathrm{m}(\mathrm{~N}-\mathrm{m}}{\mathrm{Nx}^{2}}\right]
$$

Where:
$r=$ the reliability
$\mathrm{N}=$ the number of items in the test
$\mathrm{m}=$ the mean score on the test for all the testees
$\mathrm{x}=$ the standard deviation o all the testees' score

### 2.5 Item Analysis

The items should be analyzed to determine their effectiveness. It means that the test is not finished yet once the raw mark have been obtained. ${ }^{50}$

[^19]It need further analyzed in order to get information corcerning (1) the performance of the students a group, thus informing the teacher about the effectiveness of the teaching, (2) the performance of individual student, and (3) the performance of each of the items comprising the test. ${ }^{51}$

Concerning the performance of the students as a group and individual student, item analysis shows not only the types of errors most frequently made, but also the actual reasons for the errors being made. It helps the teachers to know how effective the teaching learning activities are. For the items itself, items analysis shows which items will be used and which items will be rewritten or replaced since it tells us whether an items is too difficult or too easy, whether all the distracters function is intended, and how will it discriminate between high and low score on test.

In items analysis, all items should be examined from the point of view of (1) their difficulty level and (2) their level of discrimination. ${ }^{52}$

### 2.5.1 Index of Difficulty

The index of difficulty shows how easy or difficult the particular item proved in test. ${ }^{53}$ It expresses the percentage of the students who

[^20]answer the item correctly. In addition, Oller points out that items difficulty is about how difficult or how easy a test item for the students being investigated. ${ }^{54}$ A good test item must not too difficult or too easy for the students.

The students' score must be analyzed in order to know exactly the index of difficulty of the test. The index of difficulty is calculated by using formula below: ${ }^{55}$

## $\underline{\text { Correct } U+\text { Correct L }}$

$$
\text { F. } V=\quad 2 n
$$

Where :

$$
\begin{array}{ll}
\text { F.V } & =\text { the index of difficulty } \\
\text { Correct } \mathrm{U} & =\text { the number of students in upper } \\
& \text { group who answer the } \\
& \text { items correctly }
\end{array}
$$

[^21]
## Correct $\mathrm{L}=$ the number of students in lower group who answer the items correctly $\mathrm{n} \quad=$ the number of students in each group

The criteria to interpret the result above are as follows: ${ }^{56}$
$0,71-1,00=$ easy
$0,31-0,70=$ moderate
$0,00-0,30=$ difficult

The criteria above show that if the index of difficulty shows 1,00 , the test is too easy since the students can answer all items. It is not good to be given to the students. Moreover, if the index of difficulty shows 0,00 , the test is too difficult since the students cannot answer all the items. This test

[^22]is also not good be given. The test which is good to be given to the students is the test with criterion between $0,31-0,70 .{ }^{57}$

It is important to recognize that an item which half of the students answer correctly has the highest possible discriminating potential. Consider an item which $80 \%$ of the upper group and $20 \%$ of the lower group answer correctly. According to the rule of thumb for items answered by half or less of the students, the maximum discriminating ability of the item is 80 plus 20 , or 100 . Since the index of discrimination of the item is 60 , the discriminating efficiency is $60 \%$. As the difficulty of an item varies so that more than half of the combined upper and lower groups answer the item correctly, the discriminating ability will decrease from 100. The lower limit of the maximum discriminating ability is zero when all of the combined upper and lower groups, or none of them, answer an item correctly.

A useful rule of thumb in interpreting the index of discrimination is to compare it with the maximum possible discrimination for an item. The maximum possible discrimination is a function of item difficulty. When half or less of the sum of the upper group plus the lower group answered the item correctly, the maximum possible discrimination is the sum of the

[^23]proportions of the upper and lower groups who answered the item correctly. For example, if $30 \%$ of the upper group and $10 \%$ of the lower group answered the item correctly, the maximum possible discrimination is 30 plus 10 , or 40 .

### 2.5.2 Index of Discrimination

The discrimination index of item discriminates between the testees, separating the more able testees from the less able (Heaton, 1988: 179). In other words it can be said that the index of discrimination is the ability to differentiate between students who achieve well (upper group) and those who achieve poor (lower group). Estimate the index of discrimination is by comparing the member of students in the upper group and the lower group who answer the items correctly.

The index of discrimination can be calculated by using formula below: ${ }^{58}$

## Correct U - Correct L

$$
\mathrm{D}=\quad \mathrm{n}
$$

Where:

$$
\begin{aligned}
\mathrm{D} & =\text { the index of discrimination } \\
\text { Correct } \mathrm{U} & =\text { the number of students in upper } \\
& \text { group who answer the } \\
& \text { items correctly }
\end{aligned}
$$

[^24]Correct $\mathrm{L}=$ the number of students in lower group who answer the items correctly $\mathrm{n} \quad=$ the number of students in each group

The criteria used to interpret the result above are as follows ${ }^{59}$ :
$0,00-0,20=$ poor
$0,20-0,40=$ satisfactory
$0,40-0,70=\operatorname{good}$
$0,70-1,00=$ excellent
Discrimination indices can range from +1 ( $=$ an item which discriminates perfectly) through 0 (= an item which does not discriminate in any way at all) to -1 (= an item which discriminates in entirely the wrong way). ${ }^{60}$ It means that if the test discriminates perfectly with the index of discrimination +1 , all the students in upper group can answer all the items of the test correctly, while the students in the lower group cannot. On the other hand, if the index of discrimination is -1 , all the students in upper group cannot answer all the items correctly, but all the students in

[^25]the lower group can. This kind of test is entirely wrong and must be replaced. However, if both the students in the upper group and in the lower group can or cannot answer the items correctly, so the index of discrimination is 0 . This kind of test does not discriminate in any way at all.

### 2.6 Review of the Previous Studies

Concerning with this study, there are four studies taken before. Those studies analyzes the quality of the teacher-made English test items concerning its content validity, reliability, index of difficulty, and index of discrimination. Those studies are:

1. An analysis of the English test items of the first term of local summative test for the second year students of junior high schools in Mojokerto done by Suharman. He finds that the test does not have adequate content validity, has adequate discrimination reliability, and has acceptable facility value, does not have adequate discrimination index and have the effectiveness of distracters.
2. An analysis of the reading section of the English test items of UAN 2003/ 2004 done by Nurul Khoiriyah. She finds that the test has
high content validity, acceptable reliability, does not have acceptable index of difficulty, has poor discrimination index, and has effective distracters.

[^0]:    ${ }^{8}$ Nurkancana, Wayan and P.P.N. Sumartana. 1986. Evaluasi Pendidikan. Surabaya: Usaha Nasional.pg. 2
    ${ }^{9}$ Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. USA: Oxford University Press.pg. 18

[^1]:    ${ }^{10}$ Nurkancana, Wayan and P.P.N. Sumartana. 1986. Evaluasi Pendidikan. Surabaya: Usaha Nasional.pg. 2
    ${ }^{11}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes:
    Fundamental of Testing. USA: Harcourt Brace Javanovich.pg. 12

[^2]:    ${ }^{16}$ Brown, Douglas H. 2004. Language Assement Principles and Classroom Practices. USA: San Francisco State University.pg. 6

[^3]:    ${ }^{17}$ Grounlund, Norman E. 1976. Measurment and Evaluation in Teaching. New York: McMillan Publishing.pg. 16

[^4]:    ${ }^{18}$ Johnson, David W. And Roger T. Johnson. 2002. Meaningful Assessment: a Manageable and Cooperative Process. USA: Allyn and Bacon.pg. 7
    ${ }^{19}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGraw-Hill.pg. 1

[^5]:    ${ }^{20}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGraw-Hill.pg. 1
    ${ }^{21}$ Johnson, David W. And Roger T. Johnson. 2002. Meaningful Assessment: a Manageable and Cooperative Process. USA: Allyn and Bacon.pg. 53
    ${ }^{22}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 144

[^6]:    ${ }^{23}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGraw-Hill.pg. 1

[^7]:    ${ }^{24}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 144
    ${ }^{25}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 144
    ${ }^{26}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 25

[^8]:    ${ }^{27}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 25
    ${ }^{28}$ Johnson, David W. And Roger T. Johnson. 2002. Meaningful Assessment: a Manageable and Cooperative Process. USA: Allyn and Bacon.pg. 66
    ${ }^{29}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes:
    Fundamental of Testing. USA: Harcourt Brace Javanovich.p. 111

[^9]:    ${ }^{30}$ Khoiriyah, Nurul. 2005. An Analysis on the Reading Section of the English Test Items of UAN 2003/2004. Unpublished S-1 Thesis. Surabaya: Universitas Negeri Surabaya.

[^10]:    ${ }^{31}$ Nurgiyantoro, Burhan.1987. Penilaian dalam Pengajaran Bahasa dan Sastra. Yogyakarta: BPFE.pg. 13
    ${ }^{32}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 163
    ${ }^{33}$ Khoiriyah, Nurul. 2005. An Analysis on the Reading Section of the English Test Items of UAN 2003/2004. Unpublished S-1 Thesis. Surabaya: Universitas Negeri Surabaya.pg. 22

[^11]:    ${ }^{34}$ Nurkancana, Wayan and P.P.N. Sumartana. 1986. Evaluasi Pendidikan. Surabaya: Usaha Nasional.pg. 29

[^12]:    ${ }^{35}$ Weir, Cyril J. 1990. Communicative Language Testing. UK: Prentice Hall International.pg. 43
    ${ }^{36}$ Nurkancana, Wayan and P.P.N. Sumartana. 1986. Evaluasi Pendidikan. Surabaya: Usaha Nasional.pg. 31

[^13]:    ${ }^{37}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 28

[^14]:    ${ }^{38}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes: Fundamental of Testing. USA: Harcourt Brace Javanovich.pg. 79

[^15]:    ${ }^{39}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes: Fundamental of Testing. USA: Harcourt Brace Javanovich.pg. 81

[^16]:    ${ }^{40}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGrawHill.pg. 13
    ${ }^{41}$ Bloom, Benjamin S. T all. 1981. Evaluation to Improve Learning. USA: McGrawHill.pg. 72
    ${ }^{42}$ Grounlund, Norman E. 1976. Measurment and Evaluation in Teaching. New York: McMillan Publishing.pg. 79

[^17]:    ${ }^{43}$ Grounlund, Norman E. 1976. Measurment and Evaluation in Teaching. New York: McMillan Publishing..pg. 81
    ${ }_{45}^{44}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 160
    ${ }^{45}$ Weir, Cyril J. 1990. Communicative Language Testing. UK: Prentice Hall International.pg. 24

[^18]:    ${ }^{46}$ Oller, John W. 1979. Language Test at School. USA: Longman.pg. 4
    ${ }^{47}$ Tuckman, Bruce W. 1975. Measuring Educational Outcomes: Fundamental of Testing. USA: Harcourt Brace Javanovich.pg. 256
    ${ }^{48}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 162

[^19]:    ${ }^{49}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 164
    ${ }^{50}$ Harris, David P. 1969. Testing Language as a Second Language. USA: McGrawHill.pg. 105

[^20]:    ${ }_{52}^{51}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 178
    ${ }^{52}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 178

[^21]:    ${ }_{54}^{53}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 178
    ${ }_{55}^{54}$ Oller, John W. 1979. Language Test at School. USA: Longman.pg. 246
    ${ }^{55}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 182

[^22]:    ${ }^{56}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 212

[^23]:    ${ }^{57}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 212

[^24]:    ${ }^{58}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 182

[^25]:    ${ }^{59}$ Arikunto, Suharsimi. 1986. Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.pg. 223 ${ }^{60}$ Heaton, J.B. 1988. Writing English Language Test. New York: Longman.pg. 180

