CHAPTER II

ANALYSIS OF ENGLISH SUMMATIVE TEST FOR SECOND GRADE OF SENIOR HIGH SCHOOL

This part provides previous research and theoretical review which covers definition of evaluation, definition and types of test, categories of good test, types of test item.

A. Previous Research

The researcher has some relevant previous researches that support this research. The research is inspired from final project of:

1. "An Analysis of the English Summative Test Items in terms of Difficulty Level for the Second Year Students of MTs. Darul Ma'arif Jakarta" by Rika Amelia (2010). This research is purposed to measure the difficulty level of the English Summative Test items by calculation the student's correct response from the upper and lower group with J.B. Heaton's formula referred from his book "Writing English Language Test". Research question of this research is "Does the English Summative Test items for the second year students of MTs. Darul Ma'arif Jakarta have a good quality in terms of difficulty level?". The result of this research is interpreted by the Suharsimi Arikunto's criteria of items referred from his book "Dasar-dasar Evaluasi Pendidikan" that there are 20 items regarded as difficult level, twenty items regarded as moderate level, and nine items regarded as easy level. All of items have been counted by dividing the total of difficulty level of the items with the total number of students is 0.45. in the end of this research, the researcher has been concluded that the English Summative Test items for the second year students of MTs. Darul Ma'arif qualified as a good test seen from the difficulty level of all item which is at moderate level, because it ranges from 0.30 up to 0.70[9].

[9] Amelia, Rika (105014000357), *An Analysis of the English Summative Test Items in terms of Difficulty Level for the Second Year Students; A case study of MTs. Darul Ma'arif Jakarta*, (Jakarta: Syarif Hidayatullah State Islamic University, 2010)

2. "An Analysis on Content Validity of English Summative Test" by Moh. Cahyono Adhi Nur (2010). This research is almost same as Rika Amelia's skripsi, it has purposed to analyze the English Summative Test but this analyzes its Content validity. Research question of this research is try to find out whether the material tested in English Summative Test for the second grade of SMA Negeri 87 Jakarta in line with their KTSP?. In the end of this research, the researcher has been concluded that the total items contained by English Summative Test of the odd semester for the second grade students of SMA Negeri 87 Jakarta to their English KTSP there are 10 items measuring eight of ten function skills to be measured (8/10 x 100% = 80%). And it falls in range of 76% - 100% which means "Good".[10]

These previous researches have similarites to this research. The similarities are on the object and the method of study. They analyze the English Summative test as the object of study and they use quantitative descriptive as a method of study. The different between those previous research and this research is on the point of view analyzed of English Summative. Moh. Cahyono analyzed just focused on their content validity. While, Rika Amelia's research just focused on their difficulty level. In this research, the writer analyzes whole items of the test including their validity, reliability and practicality, and their item analysis which consists of difficulty level, discriminating power and their distractor analysis.

B. Theoretical Framework

1. The Definition of Evaluation

Evaluation is the process of making judgments about what is good or desirable. For example, judging whether a student is performing at a high enough level to move on to the next reading level or whether to carry out a

[10] Nur, Moh. Cahyono Adhi (203014001575), *An Analysis on Content Validity of English Summative Test; A Case Study at Second Grade at SMA 87 Jakarta, South Jakarta*, (Jakarta: Syarif Hidayatullah State Islamic University, 2010)

particular instructional activity requires evaluation[11]. It is accordance with what Robert stated. He said that evaluation is the process of making value judgments understates the complexity and difficulty of the effort required. When it has been determined that evaluation is needed, the evaluator must decide what kind of information is needed, how the information should be gathered, and how the information should be synthesized to support the outcome-the value judgment.[12]

From those explanations, it can be said that evaluation concerns with information gathering as it is with making decisions. In addition, the term is used to refer to the product or outcome of the process. The term evaluation usually refers to the process of making judgments, assigning value, or deciding on worth. For example, a test is an assessment technique to collect information about how much students know on a particular topic. Assigning a grade, however, is an evaluative act, because the teacher is placing a value on the information gathered on the test. So, evaluation is used to provide feedback and to generate learning with strong emphasis on student participation in self-evaluation.

2. The Definition of Test

When people hear the word assessment and evaluation, they often think of tests. The test word is derived from *testum.* It means the plate to keep off the pure metals. Test is a tool or procedures used to measure and know something based on the specific clues. If we are talking about test, there are some terminologies. Firstly is *testee*, the despondences that are doing the test. Secondly is tester, is someone who is responsible to collect and resume the despondences' result.[13]

---

[11] Airasian, Peter W., *Classroom Assesment; Concepts and Applications,* (New York: McGraw-Hills companies, inc., 2012), 7th edition, p.39

[12] Ebel, Robert L. and David A Frisbie, *Essential of Educational Measurements,* (New Delhi: Prentice Hall inc., 1991), 5th edition, p.35

[13] Arikunto, Suharsimi, *Dasar-dasar Evaluasi Pendidikan,* (Jakarta: PT. Bumi Aksara, 2005), edisi revisi, p.52-53.

Peter W. Airasian stated that a test is a formal, systematic procedure used to gather information about students' achievement or other cognitive skills[14]. Then, Glenn Fulcher added that testing is primarily about establishing *ways of making decisions* that are (hopefully) not random, and seen as 'fair' by the population[15].

In other word, a test is a method of measuring person's ability or knowledge in a given domain. it consists of some systematic procedures for gathering data about students' achievement and can be instrument, techniques, or procedures to have the students' respond through performance or tasks in the form of set question.

According to Arthur Hughes testing has several purposes:

1. To measure language proficiency
2. To discover how successful students have been in achieving the objectives of a course of study.
3. To diagnose students' strengths and weaknesses, to identify what they know and what they do not know.
4. To assist placement of students by identifying the stage or part of a teaching program most appropriate to their ability.[16]

[14] Airasian, Peter W., *Classroom Assesment; Concepts and Applications*, (New York: McGraw-Hills companies, inc., 2012), 7th edition, p.39

[15] Fulcher, Glenn, *Practical Language Testing*, (Britain: Hodder Education, An Hachette UK Company, 338 Euston Road, 2010), p.4

[16] Hughes, Arthur, *Testing for Language Teacher*, (Cambridge: Cambridge University Press, 2003), p.8

3. Types of Test

There are four types of test according to Arthur Hughes. There are:

a. Proficiency Test

According to J.B. Heaton that the proficiency test is concerned simply with measuring a student's control of the language in the light of what he or she will be expected to do with it in the future performance of a particular task.[17]. while James Dean Brown stated that a proficiency test assess the general knowledge or skill commonly required or prerequisite to entry into (or exemption from) a group of similar institution."[18]

Proficiency tests are kinds of tests designed to measure people's ability in a language, regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses that people taking the test may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient.

Proficiency tests are often used for placement or selection, and their relative merit lies in their ability to spread students out according to ability on a proficiency range within the desired area of learning.

b. Achievement Test

As its name reflected, the purpose of achievement test is to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives. H. Douglas Brown stated that

[17] Heaton, J.B., *Writing English Language Tests*, (USA: Longman inc., 1988), p.173

[18] Brown, James Dean, *Testing in Language Programs*, (New Jersey: Prentice Hall Regents, 1996), p.10

an achievement test is related directly to classroom lessons, units, or even a total curriculum[19].

According to Mehres and Lehmann stated on the book *A guide to language testing: Development, Evaluation and Research*, achievement test may be used for program evaluation as well as for certification of learned competence. It follows that such tests normally come after a program of instruction and that the components or items of the tests are drawn from the content of instruction directly.[20]

Thus it can be inferred that achievement tests are used to measure the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning program. Achievement tests are also used by teacher to motivate students to study. If students know they are going to face a quiz at the end of the week, or an end of semester achievement test, the effect is often an increase in study time near the time of the test.

According to Arthur Hughes, there are two kinds of Achievement test:[21]

1) Summative Tests (Final achievement tests)

Summative assessments**,** in contrast, are efforts to use information about students or programs after a set of instructional segments has occurred. Their purpose is to summarize how well a particular student, group of students, or teacher performed on a set of learning standards or objectives. Information obtained from summative assessments is used by

[19] Brown, H. Douglas, *Teaching by Principles; An Interactive Approach to Language Pedagogy*, (New York: A Pearson Education Company, 2001), 2nd edition, p. 391

[20] Henning, Grant, *A Guide to Language Testing : Development, Evaluation and Research*, (China: Heinle & Heinle Publisher, 2001), p.6

[21] Hughes, Arthur, *Testing for Language Teacher*, (Cambridge: Cambridge University Press, 2003), p.13

teachers to determine grades and to explain reports sent to students and their parents.[22]

In summative testing, it is expected that test scores to carry *generalizable* meaning; that is, the score can be interpreted to mean something beyond the context in which the learner is tested.[23]

It is concluded that, summative test is administered at the end of a course of study. They may be written and administered by ministries of education, official examining boards, or by member of teaching institutions. This test is designed to know how succesful students have mastered the previous materials of a long period of course.

2) Formative Test (Progress achievement tests)

This is a way of measuring progress would be repeatedly to administer final achievement tests, they are hope to increase scores indicating the progress made.

Peter W. Airasian stated that, formative tests take place while interacting with students and focused on making quick and specific decisions about what to do next in order to help students learn. They all rely on information collected through either structured formal activities or informal observations made during the process of instruction.[24]

Formative tests are typically designed to measure the extent to which students have mastered the learning outcomes of a rather limited segment of instruction, such as a unit or a textbook chapter. These tests are similar to the quizzes and unit tests that teachers have traditionally used, but they place greater emphasis on (1) measuring all of the

[22] Arends, Richard I., *Learning to teach*, (New Yoek: McGraw Hills, 2012), 9th edition, p.217

[23] Fulcher, Glenn, *Practical Language Testing*, (Britain: Hodder Education, An Hachette UK Company, 338 Euston Road, 2010), p.20

[24] Airasian, Peter W., *Classroom Assesment; Concepts and Applications*, (New York: McGraw-Hills companies, inc., 2012), 7th edition, p.99-100

intended outcomes of the unit of instruction, and (2) using the results to improve learning (rather than to assign grades).[25]

The result of formative test gives the information about how well students have mastered a particular material. The purpose is to identify the students' learning successes and failures so that adjustments in instruction and learning can be made. The formative test also determines whether a student has not been mastered the learning tasks being taught, it can be prescribed how to remedy the learning failures.

[25] Hughes, Arthur, *Testing for Language Teacher*, (Cambridge: Cambridge University Press, 2003), p.15

c. Diagnostic Test

James Dean Brown stated that a diagnostic test is designed to determine the degree to which the specific instructional objectives of the course have been accomplished.[26] J.B Heaton also stated that diagnostic test is widely used; few tests are constructed solely as diagnostic tests. Note that diagnostic testing is frequently carried out of groups of students rather for individuals.[27]

In summary, diagnostic tests are designed to diagnose a particular aspect of a language and can be used to check the students' in learning a particular element of the course. For example: it can be used at the end of a chapter in the course book or after finished one particular on lesson.

d. Placement Test

The placement test provides an invaluable aid for placing each student at the most beneficial position in the instructional sequence.[28]

The purpose of placement test according to H. Douglas Brown is to place a student into an appropriate level or section of a language curriculum or school. A placement test typically includes a sampling of material to be covered in the curriculum (that is, it has content validity), and it thereby provides an indication of the point at which the student will find a level or class to be neither too easy nor too difficult, but appropriately challenging.[29]

In summary, placement tests are intended to provide information that will help to place students at the stage or in the part of the teaching learning program that most appropriate with their abilities. Most of classroom

---

[26] Brown, James Dean, *Testing in Language Programs*, (New Jersey: Prentice Hall Regents, 1996), p.15

[27] Heaton, J.B., *Writing English Language Tests*, (USA: Longman inc., 1988), p.173

[28] Gronlund, Norman E., *Constructing Achievement Tests*, (USA: Prentice Hall Inc., 1977), p.3

[29] Brown, H. Douglas, *Teaching by Principles; An Interactive Approach to Language Pedagogy*, (New York: A Pearson Education Company, 2001), 2nd edition, p. 390-391

teacher used placement test in the pretest. So that, they can know the readiness of students to begin the instructions and place them with the proper instruction in the part of teaching learning activity.

In this research, a kind of test that will be analyzed is English summative test for second grade of senior high school made by MGMP LP Ma'arif NU of Semarang district at odd semester 2013-2014.

4. Categories of Good Test

Test as an instrument of obtaining information should have a good quality. The quality of a test will influence the result of the test itself. Once the test has a good quality, the right information will be gained and used to make accurate decision to the students' achievement.

H. Doughlas Brown stated that a well constructed test should have five main characteristics which involve validity, reliability, practicality, authenticity and washback. Validity is the degree to which the test actually measures what is intended to measure. Reliability is consistent and dependable. A practicality is means of financial limitations, time constraints, ease of administration, and scoring and interpretation. Then, authenticity is defined as a concept that is a little slippery to define, especially within the art and the science of evaluating and designing tests. Meanwhile, washback is the effect of testing in teaching and learning.[30]

a. Validity

A test has validity if it measures appropriately, what it is supposed to measure. According to Heaton, the validity of a test is the extent to which it measures what is to measures and nothing else.[31]

[30] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008),p.93

[31] Heaton, J.B., *Writing English Language Tests*, (USA: Longman inc., 1988), p.159

Validity in testing and assessment have traditionally been understood to mean discovering whether a test measures accurately what it is intended to measure or uncovering the appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure.

The view of validity presupposes that when we write a test we have an *intention* to measure something, that the 'something' is 'real', and that validity enquiry concerns finding out whether a test 'actually does measure' what is intended.[32]

Based on explanation above, the researcher concluded that, the tests can can be called valid if they measure accurately and appropriately what they intend to measure.

A standard and widely adopted classification system divides validity into the following basic types: 1) content validity, 2) criterion-related validities (predictive and concurrent), and 3) construct validity.[33]

In this research, the researcher limited the study just focuses on the content validity analysis.

1) Content Validity

Content validity is a matter of determining whether the sample is representative of the larger universe it is supposed to represent.[34] It could also be defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested.[35]

[32] Fulcher, Glen and Fred Davidson, *Language Testing and Assessment; an Advanced Resource Book*, (Canada: Routledge, 2007), p.4

[33] Gronlund, Norman E., *Constructing Achievement Tests*, (USA: Prentice Hall Inc., 1977), p.131

[34] Gronlund, Norman E., *Constructing Achievement Tests*, (USA: Prentice Hall Inc., 1977), p.132

[35] Fulcher, Glen and Fred Davidson, *Language Testing and Assessment; an Advanced Resource Book*, (Canada: Routledge, 2007), p.27

In summary, content validity is how the sample of test represents the domain of task to be measured well. Furthermore, it also important to the test makers or classroom teachers that the test cover all material that are supposed to be measured. As Gronlund stated that, we can build a test that has high content validity by identifying the subject-matter topics and behavioral outcomes to be measured.[36]

Here, the writer chooses to analyze the content of validity because she wants to know how extent the conformity between the indicators which are recommended in the teacher syllabus and the content of English Summative test which has been made by MGMP LP Ma'arif NU of Semarang district. Furthermore, the researcher also limits the study focuses on 2 skills; reading and writing, because the test is given to the students practically only contained 2 skills; reading and writing.

b. Reliability

The second criterion of a good test is reliability. Reliability is the consistency of test scores across facets of the test.[37] A consistent measurement is a necessary condition for high quality educational testing. This consistency of a test is called as reliability. Reliability is a necessary characteristic of any good test: for it to be valid at all, a test must first be reliable as a measuring instrument.[38]

It is concluded that, a reliable test is consistent and dependable. If we give the same test to the same student or matched students on two different occasions, the test should produce similar results.

[36] Gronlund, Norman E., *Constructing Achievement Tests*, (USA: Prentice Hall Inc., 1977), p.132

[37] Fulcher, Glen and Fred Davidson, *Language Testing and Assessment*; an Advanced Resource Book, (Canada: Routledge, 2007), p.15

[38] Heaton, J.B., *Writing English Language Tests,* (USA: Longman inc., 1988), p.162

c. Practicality

A test is called to have a good practicality when the tests are practical and easy to administrate. According to Anas Sudjijono stated that, a good practicality on the test means that the test should have two criterion: a). Simple, it does not require much equipment or tools that are difficult to procure. b). Complete, it comes with instructions on how to do it, the answer keys and its scoring guidance.[39]

In researcher's opinion, an evaluation procedure must meet certain practical requirements. It should be easily administrated and scored and it should provide results that can be accurately applied and interpreted by the school administer.

d. Authenticity

Authenticity is a concept that is little slippery to define, especially within the art and science of evaluating and designing tests.[40] Fulcher also stated that, authenticity is defined as the relathionship between test task characteristics, and the characteristics of tasks in the real world.[41] We can concluded that authenticity is the basis how well it replicates real life in the tasks.

In a test, authenticity may be present in the following ways:[42]

1) The language in the test is as natural as possible.
2) Items are contextualized rather than isolated.
3) Topics are meaningful (relevant, interesting) for the learner.

[39] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.97

[40] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004), p.28

[41] Fulcher, Glen and Fred Davidson, *Language Testing and Assessment; an Advanced Resource Book*, (Canada: Routledge, 2007), p.15

[42] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004), p.28

4) Some thematic organization to items is provided, such as through a story line or episode.
5) Tasks represent, or closely approximate, real-world tasks.

e. Washback

The last major principle of language testing is washback. It is the effects the tests have on instruction in terms of how students prepare for the test.[43] Fulcher stated that washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning.[44] Here, the researcher can concluded that washback is generally defined as the influence of testing on teaching and learning.

According to Cyril Weir stated that, the beneficials of washback are to occur:

1) Training teacher in the new content and metodology required for the test is essential.
2) Support in the forms of appropriate teaching materials must be freely available.[45]

Here, the researcher limited the study just focuses on analysis of content validity, reliability and practicality of the test.

5. Item Analysis

A good test should also be good at its item analysis, that is some rather simple statistical ways of checking individual items. H. Douglas Brown also stated that, "there are three main components of item analysis, they are:

[43] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004), p.28
[44] Fulcher, Glen and Fred Davidson, *Language Testing and Assessment; an Advanced Resource Book*, (Canada: Routledge, 2007), p.221
[45] Weir, Cyril J., *Language Testing and Validation,* (New York: Palgrave Macmillan, 2005), p.212

difficulty level, discriminating power and the effectiveness of the distractor."[46] Meanwhile, according to Purwanto, a good test item should have three criteria; moderate difficulty level, high discriminating power and distractor analysis which work effectively.[47]

a. Difficulty Level

A good test item should have the level of difficulty, which includes easy, moderate and difficult level. An effective and good test should have the items that belong to moderate level. The item that is too easy or difficult potentially weaken the quality of the test and the valid data of information about students' achievement will not be acquired.

To conduct an item analysis, we first arange the scored test papers or answer sheets, in order from the highest score to the lowest score. Next, we separate the papers into upper and lower groups, according to their total test scores. For large groups, we would choose the upper and lower 27 percent, while for smaller groups, we would typically choose the upper and lower one-third.[48]

Then, level of difficulty must be interpreted in the rank scale of difficulty level, as follows:[49]

Table 1
The rank scale of difficulty level

Interpretation	*P*
Difficult	$0 < p < 0.30$
Moderate	$0.30 < p < 0.70$
Easy	$0.70 < p < 1.00$

[46] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004), p.58-60

[47] Purwanto, *Evaluasi Hasil Belajar,* (Yogyakarta: Pustaka Pelajar, 2009), p. 99

[48] Bachman, Lyle F., *Statistical Analyses for Language Assesment,* (Cambridge: Cambridge University Press, 2004), p.123

[49] Arikunto, Suharsimi, *Dasar-dasar Evaluasi Pendidikan*, (Jakarta: PT. Bumi Aksara, 2006), edisi revisi, p. 210

b. Discriminating Power

Discriminating power is the extent to which an item differentiates between high and low test-takers.[50] According to Purwanto, good items test is test which have high discriminating power.[51] It can be known through or by looking at the size of the item discrimination index numbers. Item discrimination index is a number that indicates the size of discriminatory power that is owned by an item. Discriminating basically is calculated based on classification into two groups, higher and lower group.[52]

The reason for identifying these two groups is that discriminating power allows teacher to contrast the performance of the upper group students on the test with that of the lower group students. To do this, teacher or test maker can compare the number of students in the upper and lower group who answered the item correctly. For example, 6 students in the upper group and 4 students in the lower group selected the correct answer. This indicates positive discrimination since the item can differentiate between these two groups.

Then, the discriminating power must be interpreted in the rank scale of discriminating power, as follow: [53]

Table 2
The rank scale of discriminating Power

Index of discriminating power *(D)*	Classification	Interpretation
Negative	*Bad*	It has bad discriminating power

[50] Brown, H. Douglas, *Language Assesment, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004) p.59

[51] Purwanto, *Evaluasi Hasil Belajar,* (Yogyakarta: Pustaka Pelajar, 2009), p. 99

[52] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p. 387

[53] Sudijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p. 389

0.00< *D* <0.20	*Poor*	It has weak/poor discriminating power
0.20< *D* <0.40	*Satisfactory*	It has satisfactory of discriminating power
0.40< *D* <0.70	*Good*	It has good discriminating power
0.70< *D* <1.00	*High*	It has high discriminating power

c. Distractor Analysis

The last concentration of item analysis activity is the effectiveness of distractor. It is a procedure specifically related to the multiple choice item. Distractor function to divert students from the correct answer if they do not know which is corect. As James Dean Brown stated that, "The primary goal of distractor efficiency is to examine the degree to which the distractors are attracting students who do not know the correct answer".[54]

To determine whether a distractor was able to work effectively if the distractors have been choosen at least by 5% of all test participants.[55]

In conclusion, the effectiveness of distractor analysis provides the information about how succesfull a distractor has diverted students who have not studied well from the correct answer.

6. Types of Test Item

An item is the basic unit of language testing. James Dean Brown stated that the item is the smallest unit that produces distinctive and meaningful

[54] Brown, James Dean, *Testing in Language Programs,* (New Jersey: Prentice Hall Regents, 1996), p.71

[55] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.411

information on a test or rating scale.[56] The item used in classroom tests are commonly divided into two broad categories: a) The objective item and b) the essay test.

a. Objective Test

Objective test items can be used to measure a variety of knowledge outcomes. The most generally useful is the multiple-choice item, but other item types also have a place. Following simple but important rules for construction can improve the quality of objective test items.[57]

In constructing an achievement test, the test maker may choose from a variety of item types. One of them is referred to as objective item. This kind of item test can be scored objectively. Furthermore, equally competent scorers can score them independently and obtain the same results.

The objective item can be classified into two types, which are selection-type test item and supply-type test item. Here, the researcher limited the study on the selection-type test item. Because the type test used in this research is selection-type test item.

The are many kinds of selection-type test items. They are multiple choice items, true-false items and matching items. Then, the researcher just focuses on the multiple choice items.

1) Multiple Choice items

Multiple choice items are made up of an item stem, which present a problem situation, and several alternatives, which provide possible solution to the problem. The options usually of, *a*, *b*, *c* or *d*. that will be counted correct, and the distractors, which are those choices that will be counted as incorrect.

[56] Brown, James Dean, *Testing in Language Programs*, (New Jersey: Prentice Hall Regents, 1996), p.49

[57] Gronlund, Norman E., *Constructing Achievement Tests*, (USA: Prentice Hall Inc., 1977), p.34

The multiple-choices item plays such an important role in the objective testing of knowledge outcomes that it will be treated first and in considerable detail.[58]

The term options refer to collectively to all the alternative choices presented to the students and includes the correct answer and the distractors. These terms are necessary for understanding how multiple-choice items function.

Designing multiple choice item test, the test maker should be consider in some ways. There are 18 basic rules in designing multiple choice item test. They are: [59]

a) Design each item to measure an important learning outcome.

b) Present a single clearly formulated problem in the stem of the item.

c) Put the alternatives at the end of the question, not in the middle.

d) Put as much of the wording as possible in the stem.

e) Eliminate unnecessary wordiness

f) Avoid negatively worded stems. "Which of the following is not....."

g) Avoid requiring personal opinion. Other item types are more suitable for this.

h) Avoid textbook wording.

i) Do not have linked or clued items.

j) All options should be homogeneous.

k) All options should be plausible.

---

[58] Gronlund, Norman E., *Constructing Achievement* Tests, (USA: Prentice Hall Inc., 1977), p.35

[59] Merchan, Damien, *Basic Principles of Test Construction*, www.cte.cornell.edu, assessed at 14 December 2014.

l) Put repeated words in the stem, not in the options

m) Punctuation should be consistent.

n) Make all options grammatically consistent with the stem of the item.

o) List options vertically.

p) Other options logically.

q) Use the option "all of the above" sparingly.

r) Use the option "none of the above" sparingly.

Multiple choices have some advantages. Wilmar Tinambunan writes the advantages of multiple choices as follow:[60]

a) The multiple choice item can be used for subject matter content in any different level of behavior, such as ability to reason, discriminate, interpret, analyze, infer, and solve problems.

b) It has less chance for the students to guess the right answer than the true-false item does because it is followed by four or five alternatives.

c) One advantage of the multiple choice items over the true-false item is that students also know what is correct rather than only know that statement is incorrect.

The weaknesses of multiple choice items according to H Douglas Brown are:[61]

a) The technique tests only recognition knowledge

b) Guessing may have a considerable effect on test scores

[60] Tinambunan, Wilmar, *Evaluation of Students Achievement*, (Jakarta: Depdikbud, 1988), p.75

[61] Brown, H. Douglas, *Language Assesments, Principles and Classroom Practices*, (USA: Pearson Education, Inc., 2004) p.55

c) The technique severely restricts what can be tested.

d) It is very difficult to write successful items.

e) Washback may be harmful.

f) Cheating may be facilitated.

b. Essay Test

Essay questions give students the greatest opportunity to supply and construct their own responses, making them the most useful for assessing higher-level thinking processess such as analyzing, synthesizing and evaluating. The essay question is also the primary means by which teachers assess students' ability to organize, express and defend ideas. The main limitations of essays are that they are time-consuming to answer and score, and they place a premium on writing ability.[62]

Essay questions require the student to create an explicit answer that the scorer can rate without describing the basis of any rating scale or without showing his or her own version of an ideal answer.[63] On the other hand, essay tests are the best measure of students' skills in higher - order thinking and written expression.[64]

There are some basic rules in designing essay item test. They are:[65]

1. Define the behavior the student is expected to exhibit before writing the prompt.

[62] Airasian, Peter W., *Classroom Assesment; Concepts and Applications*, (New York: McGraw-Hills companies, inc., 2012), 7th edition, p.149

[63] Ebel, Robert L. and David A Frisbie, *Essential of Educational Measurements*, (New Delhi: Prentice Hall inc., 1991), 5th edition, p.188

[64] Davis, Barbara Gross, *Tools for Teaching*, (San Fransisco: John Wiley & Sons, 2009), 2nd edition, p.419

[65] Merchan, Damien, *Basic Principles of Test Construction*, www.cte.cornell.edu, assessed at 14 December 2014.

2. Ask the student to use knowledge in novel situations rather than simply recalling information.

3. Ask questions that are relatively specific and focused and which will elicit relatively brief responses.

4. If you are using many essay questions in a test, ensure reasonable coverage of the course objectives.

5. Follow the test specifications in writing prompts. Questions should cover the subject areas as well as the complexity of behaviors cited in the test blueprint. Pitch the questions at the students' level.

6. Formulate questions that present a clear task to be performed.

7. Provide ample time for answering and suggest a time limit for each question.

8. Indicate the point value for each question.

9. Word questions calling for examinee opinion on controversial matters so that they ask the examinee to geve evidence to support the opinion and evaluate the examinee's response in terms of the evidence presented rather than the opinion or position taken.

10. Require all examinees to answer the same questions; don't give optional questions.

The advantages of essay test are:[66]

1. Directly assess complex higher-level outcomes.

2. Take less time to construct.

3. Assess interrogative, holistic to outcomes.

[66] Airasian, Peter W., *Classroom Assesment; Concepts and Applications*, (New York: McGraw-Hills companies, inc., 2012), 7th edition, p.154

The disadvantages of essay test are:

1. Diffiicult and time-consuming to score.

2. Provide a deep but small sample of students' performance.

3. Bluffing and the quality of writing can influence scores.

In this research, The researcher analyzes not only multiple choice test but also analyzes essay test. There are 5 numbers in the essay test of English summative test made by MGMP LP Ma'arif NU of Semarang district that will be analyzed.

7. The Principles of Test Design

Anas Sudijono stated that there are six principles in designing the test, they are:[67]

a. The test should measure objectively instructional purposes which have been stated before.

b. The test items should be sample which represent of teaching material population in one unit or course in teaching learning process.

c. The test should be various.

d. The test should be designed to measure an important learning outcomes.

e. The test should have high reliability.

f. The test can be used as gaining useful information in revising the teaching learning method.

[67] Sudjijono, Anas, *Pengantar Evaluasi Pendidikan,* (Jakarta: PT Raja Grafindo Persada, 2008), p.97-99