CHAPTER I

INTRODUCTION

This introduction chapter presents background of the study, statement of research problems, objectives of the research, significance of the research, scope and limitation of the research, and definition of key terms.

A. Background of the Research

Evaluation is an important part of every teaching and learning experience. It gives big contribution for the teaching and it provides information about the students' progress which can be used by the teachers to manage the learning task and students. As stated by Dicksin *et al* (1992:3): "Evaluation is important for the teacher because it provides a wealth of information to use for the future direction of classroom practice, for the planning of courses and for the management of learning tasks and students". Evaluation also can be said as the process to make desirable decision toward teaching and learning based on the information that has been collected, synthesized, and reflected on. Lyle F. Bachman (1990:22) states, "Evaluation can be defined as the systematic gathering of information for the purpose of making decision".

Depending on the decision being made and the information a teacher needs in order to inform that decision, testing often contribute to the process as the implementation of evaluation. Surely, a test is one kind of evaluation

instrument to collect data. Nitko (1983: 6) stated, "A test is defined as a systematic procedure for observing and describing one or more characteristics of a person with the aid of either a numerical scale or category system". In other word, a test measures a person's ability or knowledge with a number of tasks or questions. According to Henning (1987:1) "Tests in general is to pinpoint strengths and weakness in the learned abilities of students". Teachers need to do the test because through the test they are able to find out the students' achievement in mastering the lessons that have been taught and to evaluate the effectiveness of the method used and the teaching material. Valette (1977:5) states, "Through tests the teacher can evaluate the effectiveness of a new teaching method of different approach to a difficult pattern, or new teaching".

To measure the students' learning progress in the class, a teacher usually administers two kinds of test, there are formative test and summative test. The formative test is held earlier than summative test, which is held at the end of semester. Through those test the teacher can measure the students' achievement level and the degree of how far the instructional objectives can be reached by them. That reason was shown as Gronlund (1976:18) states:

> "Formative test is used to monitor learning progress during instruction. Its purpose to provide continuous feedback to both pupil and teacher concerning learning successes and failures and summative test typically comes at the end of a course 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 pupil mastery of the intended learning outcomes"

To get accurate measure a test must have a good quality, because a good test does not only influence to the students' learning but also influence to the teacher in order to improve teaching and learning process. A test is said a good test, if it has to fulfill the characteristics of good test; validity, reliability and practically. Harris (1963:13) states, "All good test posses three qualities: validity, reliability, and practicality". A test can be valid if it can measure what is supposed to measure. It can be reliable if the result of test is the same when the test is administered to the same level students in the next time. In addition, it can be practical if it is easy to administer. Brown (2001:385) states as follows:

> "How do you know if a test is a "good" test or not? Is administrable within given constraints? Is it dependable? Does it accurately measure what you want it to measure? These answer can be answered through three classic criteria for "testing a test": practicality, reliability, and validity".

The problem, which is often forgotten by teachers, is the follow up of test implementation related to the test item itself. In fact, they do not criticize whether the test fulfilled the criteria above or not. Whereas, it really required an analysis of test items namely "*item analysis"*. Through analyzing test item, the teacher can identify good item and poor item and differ between the students who have done test either well or poorly.

According to Ahman and Glock (1976:184) the purpose of test item analysis is re-examining each test item to discover its strengths and flaws is known as item analysis. Purwanto (1991:118) also states that the main purpose of item analysis is to find out what and why test items are called as good test items and bad test items.

There are three characteristics which are usually determined for a test and it can be found by analyzing it; firstly item difficulty which indicates how difficult or easy items. Bahman (2004:125) states, "Item difficulty is defined as the proportion of test takers who answered the item correctly, and the item difficulty index value can be calculated on the basis of test takers response to the item"

Second, discrimination power which tells how well the items in separating the higher students to lower students. According to Ahman (1976:187), "The discriminating power of test item is an index that shows its ability to differentiate between pupils who have achieved well and those who have achieved poorly".

Third, item distractor for multiple-choice items, it indicates how effective each option for items. Bailey (1998:134) states, "One important aspect affecting the difficulty of multiple choice test items is the quality of distractor. Some distractor in fact might not be distracting at all and therefore serve no purpose" So it can be concluded that item analysis provide us the data whether the test item is too difficult or too easy, whether test item can

discriminate the students or not and whether all the options functioned as the examiner intended.

MAN Tulungagung 1 is one of school based Islamic school, which considered as the favorite Islamic school in Tulungagung. This school always makes evaluation that commonly carried out in form of summative test per semester. The English summative test of MAN Tulungagung 1 was held on Tuesday December 3, 2013. The item of English summative test for second year of odd semester, which has been carried out, is never analyzed before. It means that the quality of the test items was never known. Most of teacher said that they have not already analyzed the test because the time, accuracy and patience are certainly needed for doing an item analysis.

Considering this fact, the researcher was interested to analyze summative test under title **"AN ITEM ANALYSIS ON ENGLISH SUMMATIVE TEST FOR SECOND GRADE STUDENTS OF MAN TULUNGAGUNG 1 IN ACADEMIC YEAR 2013/2014"**

B. Statement of the Research Problems

Here are the research problems formulated by the researcher in which the researcher tries to answer by this research. In accordance with the background of the study, the main problems in study are formulated as follows:

1. How is the validity of English summative test for second grade students in MAN Tulungagung 1?
2. How is the reliability of English Summative test for second grade students in MAN Tulungagung 1?
3. How is the level of difficulty of English Summative test for second grade students in MAN Tulungagung 1?
4. How is the discriminating power of English Summative test for second grade students in MAN Tulungagung 1?
5. How is the effectiveness for each distractor of English Summative test for second grade students in MAN Tulungagung 1?

C. Objective of the Research

According to the research problems that are previously defined, the purposes of this research are to describe information about the English Summative test of MAN Tulungagung 1 in academic year 2013/2014, which cover:

1. The validity
2. The reliability
3. The level of difficulty
4. The discriminating power
5. The effectiveness for each distractor

D. Significance of the Research

Firstly, it provides the information to the researcher especially, and English teacher/institution of how to analyze test items in terms of validity, reliability difficulty level, discrimination, and effectiveness for each distractor.

Secondly, it informs the English teacher, institution or test designer about the quality of test items in term validity, reliability, difficulty level, discrimination, and effectiveness for each distractor. Through this research, the teacher, institution or test designer can also know the good items the students' or candidates' achievement in mastering the materials that to be taught.

E. Scope and Limitation of The Research

The scope of this research covers validity, reliability difficulty level, discrimination, and effectiveness for each distractor of the English summative test for second grade students in MAN Tulungagung 1 in academic year 2013/2014. The research has analysis limitation; the validity only content validity and construct validity. The researcher also cannot guarantee whether the students cheated or not to answer every item.

F. Definition of Key Terms

1. Item Analysis

Nitko (1983:284) "item analysis refers to the process of collecting, summarizing, and using information about pupils' responses to items"

2. Validity

Heaton (1988:159) states, "The validity of test is the extent to which it measures what is to measure and nothing else"

3. Reliability

Ahmann and Glock (1976:311) state " Reliabilty means consistency of results. This is equivalent to saying that a highly reliable instrument can be used repeatedly in an unchanging situation and produce constant or near constant results.

4. Summative Test

Vallete (1977:6) states "The summative test, is usually given at the end of a marking period and measured the sum total of the material covered"

CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter presents any reviews of related literature including the definition of test, types of test, language testing, test techniques and testing overall language ability, testing language skills and language components, criteria of a good test and item analysis.

A. The Definition of Test

A test is composed of a number of tasks or questions for students to respond. By analyzing the responses, the teacher can measure the student's achievement in the teaching learning process. Bachman (1990:20) states that, "A test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual". While Djiwandono (2008:12) states that, a test is a tool or procedure used to measure the students' language proficiency. From a test teacher will get quantitative score, which can be analyzed by them.

From those views of test, it can be concluded that a test can be instrument, techniques, or procedures to have the students' responses through tasks or performance in the form of set of questions that must be answered in order to achieve the teaching-learning objectives. In short, a test is a measurement instrument designed to assess a specific sample of individuals' behavior.

B. Types of Test

There are many kinds of tests used to measure students' achievement that can be used in an evaluation process. The type of test can be classified into two types, namely function of the test and its way of scoring.

1. Types of Test Based on Its Function

According to Andrew (1983:6), the types of functional test can be categorized into four types: placement test, diagnostic test, achievement test, and proficiency test.

a. The Placement Test

Placement test is used to place a student to appropriate level or section of a language curriculum or school. It usually happens in the beginning of course. According to Hughes (1989:14); "A placement tests are intended to provide information which will help to place students at the stage of the teaching program most appropriate to their abilities. Typically they are used to assign students to classes at different level"

b. The Diagnostic Test

Heaton (1988:173) states 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.

c. Achievement Test

These tests are used to know what students have actually learnt or on what have actually been taught. Hughes (1989:13) states that the purpose of achievement test as its name reflect is to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives. At the implementation level, the achievement test appears in two purposeful tests, they are formative test and summative test.

1) Formative Test

The teacher administers formative test during the learning progress with the aim of using the result to improve instruction and to provide continuous feedback to both students and teacher. Valette (1977:6) states "The formative test is given during the course of instruction; its purpose is to show which aspects of the chapter the student has mastered and where remedial work is necessary".

2) Summative Test

Summative test is a test that usually administered at the end of the course. Valette (1977:6) states" The summative test, on the other hand, is usually gives at the end of a marking period and measures the "sum" total of the material covered. On this type of a test, students are usually ranked and graded". Moreover, summative test is given periodically to determine at a particular

point in time what students know and do not know. Summative test at the district/classroom level is an accountability measure that is generally used as part of the grading process. Hughes(1989:11) states that "the content of summative test should be based directly on a detailed course syllabus or on the books and other material used".

d. Proficiency Test

According to Heaton (1988:172) "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 ".

2. Types of Test Based on its Way of Scoring

Based on the manner of scoring, the type of test item is divided into two general types: objective and subjective test. Heaton (1988:25) states "Subjective and objective test are terms used to refer to the scoring of tests".

a. Objective Test

An objective test item is any test item that there is only a single correct answer. In this test, the students must select one option from some alternatives. According to Valette (1977:6); "An objective test item is any item for which there is a single predictable correct answer".

b. Subjective Test Item

Subjective test is a test where in its scoring requires judgment and evaluation of scores. While Vallette (1977:6) states

> "Subjective item is one that does not have a single right answer". It means that the scoring is inconsistent and the answer of the question is in form of composition where the students are given a chance to relate their idea or argument in their own words.

C. Language Testing

Language testing will lead the teacher to know the students' improvement in learning language .According to Heaton (1988:4)

> "A language test which seeks to find out what candidates can do with language provides a focus for purposeful, everyday communication activities. Such a test will have a more useful effect on the learning of a particular language than a mechanical test of structure".

Language testing also gives backwash between the teacher and the students, as Hughes (1989:1) states "The effect of testing language on teaching and learning is known as backwash. In short, administer the language testing will be useful to teacher to get the information about the students achievement.

D. Test techniques and testing Overall Language ability

According to Hughes (1989:59) "test technique are means of electing behavior from the students which will tell the teacher about their language abilities". There are some techniques as suggested by Hughes:

1. Multiple Choice

Multiple choice items take many forms, but their basic structure is as follows

There is a *stem*

End has been her_____ half an hour

And a number of *options,* one of which is correct, the others being *distracters*:

A. During
B. For
C. While
D. Since

The advantages of multiple choice test technique are perfectly reliable, rapid, economical and open ended format. The disadvantages are giving chance the students either cheating or guessing, it is extremely difficult to make.

The multiple choice is best suited to relatively infrequent testing of large number of test takers. But, actually it will not be the best way for the students to improve their command of language of language because usually much attention is paid to improve student's guessing rather than to the content items . in addition Hughes (1986:61) consider multiple choice tests having harmful backwash.

2. Cloze, C-Test, and Dictation: Measuring overall Ability

Cloze test and dictation test technique are recommended as means of measuring ability. Cloze test is the technique, which are deleted from a text allowing few sentences of introduction. The deletions are mechanically set, usually between every fifth and eleventh word. Braley (1978:2) commented, "up to now, in the main, the results of research with the cloze test have been extremely encouraging. They have shown high validity, high reliability, objectivity, discrimination and so on, she quoted Brown (1979:13) "as

demonstrated in this and other studies, it can be valid and reliable test of overall second language proficiency". The C-test is variety of cloze test, which is considered superior to the general cloze procedure. Instead of whole words, it is the second half every second word, which is deleted. The supposed advantages of C-test technique over the more traditional one are that only exact scoring necessary and that shorter passage. The disadvantage is harder to read than cloze procedure

Dictation is technique in which the passage is read aloud to students, with pauses during which they have to write down what they heard as accurately as possible. The advantages of dictation technique are easy to create and administer. The disadvantage is easy to score.

E. Testing Language Skills and Language Components

1. Testing Listening

An effective way of developing the listening skill is trough the provision of carefully selected practical material. Such material is in many ways to that used for testing listening comprehension.

According to Hughes (1989:134), the testing listening involves listening macro skills and micro skills. The macro skills of listening include; listening for specific information, obtaining gist of what is being said, following instruction. The micro skills of listening include level interpretation of intonation patterns and recognition of function of structures. At lowest level are abilities like being able to distinguish between phonemes (for example between /w/ and /v/).

Weir (1990:57) suggested the techniques that are possibly used in testing listening;

a. Multiple choice

   This technique has some advantage and disadvantage as explained above For listening test, the problem is greater because the test takers should listen to passage while reading the alternatives.

b. Information transfer technique

   This technique is useful in testing listening since it makes minimal demands on productive skills. It can involve such activities as the labeling of diagrams or pictures, completing forms and so on

c. Dictation

   This involve the students listening to dictated material which incorporates oral message typical of those might encounter in the target situations

d. Listening recall

   The student is given printed copy of passage from which certain content words have omitted.

e. Note taking

   Where the ability to take notes while students listening to lecture are in question, this activity can be suite realistically replicated in the testing situations.

f. Recording and live presentation

The great advantage of using recordings when administrating of listening test is that there is uniformity in what is presented to the test takers.

2. Testing Speaking

Huges (1989:101) states that the objective of teaching spoken language is the development of the ability to interact successfully in that language, and that this involves comprehension as well as production. Consequently, test should elicit behavior which truly represent the students' ability and which can be scored validly and reliably.

The operation is to take part in speaking test, which may involve the following functions:

Expressing: Thanks, requirements, opinion, comment, attitude, confirmation, apology, want /need, information, complaints, reason/justification.

Narrating : sequence of events

Eliciting: information, direction, service, clarification, help, permission.

Directing: ordering, instructing, persuading, advising, warning.

Reporting: description, comment, decision.

Here are the lists of the more useful and potentially valid techniques for testing speaking ability suggested by Weir (1990:74-80):

a. Verbal essay

   The student is asked to speak for three minutes or either one or more specified general topics.

b. Oral presentation

   The student is asked to give a short talk on a topic which he has either been asked to prepare before hand or has been informed of shortly before test.

c. The free interview

   In this type of interview, the conversation unfolds in an unstructured fashion and no set of procedures is laid down in advance.

d. The controlled interview

   In this procedure, there are normally a set of procedures determined in advance for eliciting performance.

e. Information transfer: description of a picture sequence

   The students see a panel of a picture depicting a chronologically ordered sequence of events and have to tell the story in past tense. Time is allowed at the beginning for students to study the picture

f. Information transfer : question on a single picture

   The examiner asks the students a number questions about content of picture, which he has had time to study.

g. Interaction tasks

Students work in pairs and each given part of the information necessary for completion the task.

h. Role play

The student is expected to play one of the roles in an interaction which might be reasonably expected of him in the real world.

i. Imitation

The students hear a series of sentence, each of which they have repeat in turn.

Technique not recommended by Hughes (1989:119-110):

1. Prepared monologue

Some examinations require students to present a monologue on a topic after being given a few minute to prepare. If a task were carried out in the native language, there would almost certainly be considerable differences between students. For this reason and because leaving the student alone to prepare a monologue must create stress, this technique is not recommended

2. Reading aloud

There will be significant difference in native speaker performance and inevitable inference between the reading and the speaking skills

3. Reading blank dialogue

Allison (1999:119-110) added reading blank dialogue are not recommended technique. It is an indirect and artificially constrained

way of testing whether learners are able to respond meaningfully to what someone says to them. This technique is not recommended because while the students may anticipate and perhaps try to influence what another speaker next, the form and context student's own contribution are not constrained by a predetermined text that follow it. In contrast the other student's turn are already provided, which sets additional limits on what we can "say" in the context of this activity

3. Testing Reading

Reading is a receptive skill. The task of language tester is then to set reading tasks, which result in behavior that will demonstrate their successful completion. In spite of the wide range of reading material specially written adapted for English learning proposes, there are few comprehensive systematic programmers, which have been constructed from a detailed analysis of the skills required for efficient reading. Few language teachers would argue against the importance of reading; what is still urgently required in many classroom tests is greater awareness of the actual processes involved in reading and the production appropriate exercise and test materials to assist in the mastery of these processes.

Hughes (1989:116-117) states the macro skills directly related either needs or to course objectives:

- scanning text to locate specific information
- skimming text to obtain the gust
- identifying stages to an argument

- identifying examples presented in support of an argument

The micro skills underlying reading skills are:

- identifying referents of pronouns, etc
- using context to guess meaning and unfamiliar words
- Understanding relation between part of text by recognizing indicators in discourse, especially for the introduction, development, transition and conclusion of ideas

Then there is what would be recognized as the exercise of straight forward grammatical and lexical abilities, such as:

- Recognizing the significance of the use of the present continuous with future time adverbials
- Knowing that the word "brother" refers to male sibling

Weir (1990: 43-50) suggested the technique that might be used to test reading as follows:

a. Multiple choice questions (MCQs)

   It is usually set out in such way that the student is required to select the answer from the number of given option, only one of which is correct.

b. Short answer question

   This question requires the students to write down specific answers in space provided on the question paper.

c. Cloze

In the cloze procedure, words are deleted from a text after allowing a few sentences of introduction.

d. Selective deletion gap filling

In this technique the constructor should use a "rational cloze" selecting items for deletion based upon what is known about language

e. C- tests

In the C-test, every second word in a text is partially deleted. In attempt to ensure solution, students are given the first half of the deleted words. The student completes the word o the test paper and an exact word scoring procedure is adapted.

f. Cloze elide

A technique, which is generating interest, recently is where words, which do not belong, are inserted into a reading passage and students have to indicate where these insertions have been made.

g. Information transfer

One way to minimize demands on writing by test takers is to require them to show successful completion of reading task by supplying simple information in a table, following route on map, labeling picture etc

Hughes (1989:126-129) added more techniques in testing reading:

h. Identifying order or events, topics or arguments

The students can be required to number the events etc.

i. Identifying referents

One of micro-skills listed above ability to identify referents.

An example of an item to test is :

What does the word 'it' refers to…….

j. Guessing the meaning of unfamiliar words from context

This is another of the micro-skills mentioned above. Items may take the form

Find a single word in the passage (between line 1 and 25) which has the same meaning as "making of laws" (the word in the passage may have an ending like –s, -ing, -ed etc)

Hughes (1989:131) advised to obtain reliable scoring, error of grammar, spelling or punctuation should not be penalized, and if it is clear, the student has successfully performed the reading task, which the item set.

4. Testing Writing

The best way to test students' writing is to get them to write directly. Therefore, indirect testing of writing ability cannot possibly constructed as accurately as possible even by professional institutions.

According to Madsen (1989: 101), there are many kinds of writing test. The reason for this is simple; a wide variety of writing tests is needed to test the many kinds of writing tests that we engaged in. Another reason for the variety of writing tests in use is the great number of factors that can be evaluated; mechanics (including spelling and punctuation), vocabulary, grammar , appropriate content, diction (word selection),rhetorical matters

of various kinds (organization, cohesion, unity; appropriateness to the audience, topic, occasion) etc.

Weir (1990:59-66) suggested the techniques to testing writing as follows;

a. Editing task

In editing task the student is given a text containing a number of errors of grammar, spelling and punctuation of the type noted as common by remedial teachers of the students in the target group and is asked to rewrite the passage marking all the necessary corrections.

b. The direct testing of writing

With a more integrative and direct approach to the testing of writing, the tester can incorporate items which students' ability to perform certain functional tasks required in the performance of duties in the target situation ,here are some kinds of direct writing test ;

(1) essay test

This is a traditional method for getting students to produce a sample of connected writing. The stimulus is normally written and can vary in length from limited number of words to several sentence.

(2) controlled writing tasks

It tests important skills, which no other form assessment can be sampled adequately. To omit a writing task in

situations where writing tasks are an important feature of the student's real life needs might severely lower the validity of testing programs.

Hughes (1989:75) suggested three things that the tester should consider to develop a good writing test as follows:

1. Tester has to set writing tasks that are properly representative of the population of tasks that tester expect the students to be able to perform
2. The tasks should elicit samples of writing which truly represent the students' ability
3. It is essential that the samples of writing can and will be scored reliably

5. Testing Grammar

The specification of grammar test should be in line with the teaching syllabus if the syllabus lists the grammatical structures to be taught. When there is no such list, it becomes necessary to infer from text books or other teaching materials.

Heaton (1988:34-50) suggested the techniques of testing grammar as follows

a. Multiple choice items

The type multiple-choice item favored by many constructors of grammar tests incomplete statement type, with a choice of four or five options

b. Changing words

This type of item is useful for testing the students' ability to use correct tenses and verb forms

c. Broke sentence items

This type item tests the student's ability to write full sentences from series of words and phrases.

d. Constructing pairing and matching items

This type of item usually consists of a short conversation.

6. Testing Vocabulary

The specification for vocabulary achievement test should be based on all items presented to the students in vocabulary teaching. When placement test is applied the vocabulary being tested should refer to one common published word lists.

The techniques to the testing vocabulary are :

a. Picture

The use of picture can limit the students to lexical items that we have in mind

b. Definition

This may work fir a range lexical items. The following is an example of such test.

A…… is a person who looks after our teeth
…….is a frozen water

c. Gap filling

This can take the form of one or more sentence with single word missing

7. Testing Pronunciation

There are at least three techniques in testing pronunciation:

1. Pronouncing words in isolation

   The importance of listening in almost all test of speaking especially those pronunciation, should never underestimated.

2. Pronouncing words in sentences

   Students can also be asked to read aloud sentences containing the problematic sounds which want to test

3. Reading aloud

   Reading aloud can offer a useful way of testing pronunciation provided that we give a student a few minutes to look at the reading text first.

F. Criteria of a good test

There are many considerations entering into the evaluation of a test, which referred as a good test because a good test can provide available information for a good evaluation in order to measure student's comprehension of the instructional objectives, but the writer consider them under three main headings;. These are respectively validity, reliability, and practically. Validity refers to the extent to which a test measures what we actually wish to measure. According to Brown "Validity is the degree to which the test actually measures what is intended to measure…..Reliability is consistent and dependable…….And practically is means of financial limitations, time constraints, ease of administration, and scoring and interpretation".

1. Validity

The single most important characteristic of a good test is its ability to help the teacher make a correct decision of what is intended to measure. This characteristic is called *validity*. Heaton (1988:159) stated that Validity is concerned with whether the information being gathered is relevant to the decision that needs to be made. A test has validity if it measures appropriately, what it is supposed to measure.

Based on the definition, the researcher can conclude that validity of test is important to know whether a test has a good quality in testing someone's capacity. As the validity is one of the most important characteristic of test scores, the constructor of the test should know the various aspects from the validity itself and various procedures by which they are determined.

According to Heaton, a validity of a test can be seen from some aspects mentioned below:

a. Face Validity

A test has face validity if the test has a good "face" or the way the test looks. According to Heaton (1988:159): "if a test items looks right to other testers, teachers, moderators, and testers, it can be described as having at least face validity".

b. Content Validity

A test has content validity if the test contains materials that the student has been taught. To fulfill this, the teacher also should refer to the

instructional objectives of the teaching learning process.

c. Construct Validity

A test is said to have a construct validity if it can demonstrates that it measures just the ability, which it is supposed to measure .according to Heaton; "if a test has construct validity, it is capable of measuring certain specific characteristics in accordance with a theory of language behavior and learning".

d. Empirical Validity

A fourth type of validity is usually referred to as statistical or empirical validity. This validity is obtained as a result of comparing the result of the test with the result of some criterion measure.

2. Reliability

The second criterion of a good test is reliability. Robert (1961:127) Reliability has to do with the accuracy and precision of a measurement procedure. Indices of reliability give an indication of the extent to which a particular measurement is consistent and reproducible. There are some ways to get reliability coefficient;

a. Test retest method

To arrive at reliability coefficient of a test, first , the teacher have to get two sets scores and for comparison. The most obvious way of obtaining these is to get a group subject to take the same test twice. Person –Product moment is usually used to find correlation.

Formula Person Product moment

$$\mathbf{r_{xy}} = \frac{\mathbf{N\sum xy - (\sum x)\,(\sum y)}}{\sqrt{[\mathbf{N\sum x^2 - (\sum x)^2}][\mathbf{N\sum y^2 - (\sum y)^2}]}}$$

Notes,

r = Pearson product-moment reliability coefficient

N= total respondent

x = variable x

y= variable y

∑xy= the total of x and y

b. Split half method

The subject takes the test in the usual way but each subject is given two scores. One score is for one-half of the test, the second is for the other half. After that, two sets of score are calculated by using Pearson product moment, then to get coefficient using Spearman-Brown Prophecy Formula:

$$\boldsymbol{r_{11}} = \frac{\mathbf{2r}}{\mathbf{1+}r}$$

c. Kuder-richardson reliability

It requires test administration only once. One correct answer is one while incorrect answer is 0. There are calculated by 2 formulas; KR-20 and KR-21

KR-20$=\frac{k\,[1-\sum pq]}{k-1 \quad s^2}$ **Kr-21** $= \frac{k\,[1-X(k-X)}{k-1 \quad ks^2}$

3. Practicality

Practicality is concerned with a wide range of factors economy, convenience and interpretability that determine whether a test is practical for widespread use. Stanley (1964:311) "Practically is concerned with a wide range of factors economy, convenience, and interpretability that determine whether a test is practical for widespread use".

A test maybe a highly reliable and valid instrument but still is beyond our means facilities. The teacher or someone who makes the test should keep in mind a number of very practical considerations. There are many factors of practicality; economy, scorability, and administrability.

G. Item Analysis

After a test has been administered and scored it is usually desirable to evaluate the effectiveness of the items. This is done by studying the students' responses to each item. When formalized, the procedure is called item analysis. Nitko (1983:342) states, "Item analysis refers to the process of collecting, summarizing, and using information about pupils' responses to items". Meanwhile Madsen (1983:180) explains:

> "The selection of appropriate language items is not enough by itself to ensure a good test. Each questions needs to function properly; otherwise, it can weaken the exam. Fortunately, there are some rather some simple statistical ways of checking individual item. This procedure is called 'item analysis'."

An item analysis also is a systematic procedure which provides some information about the quality of the test item, concerning each of the following points:

1. The difficulty of the item
2. The discriminating power of the item
3. The effectiveness of each alternatives or distracters.

Thus, item analysis information can tell the evaluator or constructor if an item was too easy or too hard, how well it discriminated between high and low scorers on the test, and whether all of the alternatives functioned as intended. According to Suharsimi Arikunto(1987:205) the aims of item analysis are to help the evaluator to identify bad test items, getting the information about test items in order to be improved later and describe the quality of test that the evaluator made.

Item analysis has several benefits. First, it provides useful information for class discussion of test. Second, it provides data for helping the students improve their learning. Third, it provides insights and skills which lead to the preparation of better tests on future occasions.

Finally, the researcher concludes that item analysis is very important to do in order to get information of the quality of the test item, whether it is good item or poor item.

1. Difficulty Level of the Item

The difficulty level of item means the percentage of pupils who answer correctly each test item. "The item difficulty is fraction of the

persons taking an item who answer it correctly". Heaton (1988:178) states:

> "The index of difficulty "(of facility value) of an item simply shows how easy or difficult the particular item provide in the test. The index of difficulty (facility value) is generally expressed as the fraction (percentage) of the students who answered the item correctly".

A good test item should have a certain degree of difficulty. It may not be too easy or too difficult because the test that is too easy or too difficult will yield same score distribution that make it hard to identify reliable differences in achievement between the pupils who have done well and these who have done poorly. Arikunto (2012:222) says the good test item means the test item which is neither too easy nor too difficult. The easy test item cannot stimulate the students to answer the question. In other hand, the difficult test items cannot support the students to answer because of out of weight.

By analyzing the students' response to the items, the level of difficulty of each item can be known and the information will be helpful for teacher in identifying concepts to re-teach the study material. In addition, by analyzing the facility value, the teacher will know if the item is easy, moderate, or difficult, Thoha (2003:145) states that in term of level of difficulty, good test item is neither too difficult nor too easy because it will affect to the discrimination power. Both high and low difficulty level lead not to have good discrimination power. To measure the difficulty level of each item, the writer uses the formula stated in Heaton's book:

$$P = \frac{NP}{P}$$

Where,

P= level of difficulty

NP= the right response

N= the number of student

2. The Discriminating Power of Item

According to Fernandes (1984:27) states that Discrimination Power of a test is ability to separate good students from poor students. These groups are defined by their scores on the test whole. The difference between percentage of the top scoring 27% of students get the item right in its discrimination index.

To know item discrimination is separating the highest scoring group and the lowest scoring group from the entire sample based on total score on the test. The students with highest total scores are compared in their performance with the students with lowest total scores using the formula:

D= PA-PB

Where,

D= Discrimination power

PA= Proportion of higher group

PB= Proportion of lower group

3. The effectiveness of each alternatives or distractors.

It is very important to do distractor analysis in order to know whether the distractors provided are useful or not. Particularly for multiple choice item include the stem (the main part of the item at the top) and the options (which are the options will be counted as incorrect). The primary goal of distractor efficiency analysis is to examine the degree to which the distracters are attracting the students who do not know the correct answer. To do this for an item, the percentages of students who choose each option are analyzed. If this analysis can also give the percentages choosing each option in the upper, middle and lower groups, the information will be even more interesting and useful.

CHAPTER III

RESEARCH METHOD

This chapter presents research design, population and sample of study, research instrument, data collection method and data analysis.

A. Research Design

Research design is the process that guides researchers on how to collect, analyze and interpret data. The researcher tried to describe the quality of summative test by analyzing the test items, so that the research design used in this research was descriptive quantitative approach, in reason of that the analysis will be dealing with number as well percentage

Cohen (2007:205) states that descriptive research looks at individuals, groups, institutions, methods and materials in order to describe, compare, contrast, classify, analyze and interpret the entities and the events that constitute their various fields of inquire.

B. Population and Sample

Population is the group to which researcher would like the result of the research to be generalized. Ary *et al* (2002:138) state, "Population is defined as all members of any well-defined class of people, events or objects". In this research, the population referred to items of English summative test in second grade students of MAN Tulungagung 1, which consist of 50 test items.

Sample is small group that is observed. According to Ary *et al* (2002:138) "Sample is a part of population, which wants to be analyzed". The researcher took the population as the sample because the population of this researcher was only 50 test items. In other word, the number of population and sample here were same.

C. Research Instrument

The term instrument used here refers to many kinds of tools employed by the researcher to obtain information. Fraenkel (2005:112) states: "Instrument is the device the researcher uses to collect data". The instrument of this research was document in form the summative test. Besides that, the syllabus and theory of language testing were also used as the basis summative test analysis.

D. Data Collecting Method

The data collecting method is needed to get data in the research. Nazir (1988:211) states, "Collecting data is a systematic and standard procedure to get data needed". The researcher used the documentation as the technique collecting method because the data was in form of document. According to Tanzeh (2011:93), "Documentation is collecting data by looking or writing a report that available such as written material or film". The data collecting method and research instrument in this research were same, that was documentation.

Documents can be classified into three categories; Bikken (1998:58) states the three categories of document as follows:

a. Personal documents: those produced by individuals for private purposes and limited use such as letters, diaries, autobiographies, family photo albums and other visual recording.

b. Official documents : produced by organizational employees for record-keeping and dissemination purposes such as memos, files, yearbooks and the like are used to study bureaucratic

c. Popular culture document: these are produced for commercial purpose to entertain, persuade, and enlighten the public such as commercial, TV programs, news reports, or audio and visual recording.

The researcher used the official documents, the documents used by the researcher were: 1) The syllabus of first semester, 2) The answer sheet of the summative test, 3) The key answer sheet, 4) The summative test

E. Data Analysis

In this study, the researcher used quantitative method to conduct item analysis. There are five points of analysis covering:

1. The Validity

First, the researcher used content validity to see how well the content of test represents the entire universe of content, which might be measured. As the name implies, content validity is concerned with whether or not the content of the test is sufficiently representative and

comprehensive for the test to be a valid measure of what it is supposed to measure that can be best examined.

To know whether the test has good content validity, the researcher used the syllabus to get the clear specification of the skills or components that it is meant to cover, then compared the test specification and test content. Finally, the researcher gave the percentage of skills being tested based on the specification provided.

Second, the researcher analyzed the construct validity. A test, part, or testing technique is said to have construct validity if it can be demonstrated that it measures just the ability, which is supposed the measure. The word "construct" refers to any underlying ability which is hypothesized in a theory of language ability, so that the researcher provided the techniques of test used then connected those to the theory of language testing to know whether the test has good construct validity or not.

2. The Reliability

The researcher used KR-20 formula to compute the reliability of test as follows:

$$\mathbf{r_{11}} = \left(\frac{n}{n-1}\right)\left(\frac{s_t^2 - p_1 q_1}{s_t^2}\right)$$

Note:

$\mathbf{r_{11}}$ = reliability coefficient

n = number of test items

S_t^2 = standard deviation

p_1 = the right responds

q_1 = the wrong respond

After using the KR-20 formula, the researcher classified the reliability coefficient which taken from Sudjiono (1996:388), as the table follows:

Table 3.1 The Classification of Reliability Test

Reliability test coefficient	Classification
0.99-1.00	More highly
0.77-0.89	High
0.50-0.69	Fair
0.30-0.49	Low
<0.30	Very low

3. The Level of Difficulty

The formula for item difficulty is:

$$P = \frac{NP}{N}$$

Where,

P= level of difficulty

NP= the right response

N= the number of student

To know the classifications of the difficulty level, the researcher used the classification referred to Arikunto (2012:210), the following is the classification and interpretation of difficulty level:

Table 3.2 Classifications of Difficulty Indices

Difficulty Level	Classification
0.00-0.30	Difficult
0.30-0.70	Fair
0.70-1.00	Easy

4. The Discrimination Power

The first step of computing item discrimination was separating the highest scoring group and the lowest scoring group from the entire sample on the basis of total score on the test. The students with highest total scores were compared in their performance with the students with lowest total scores using the formula:

D= PA-PB

Where,

D= Discrimination power

PA= Proportion of higher group

PB= Proportion of lower group

The proportion of the higher group (PA) can be obtained through the following formula:

$$PA=\frac{BA}{JA}$$

Where,

PA= the proportion of the higher group

BA =the number of correct responses in the higher group

JA= the number of testers in the higher group

Meanwhile, the proportion of the lower group (PB) can be obtained through the following formula:

$$PB=\frac{BB}{JB}$$

Where,

PB = the proportion of the higher group

BB= the number of correct responses in the higher group

JB= the number of testers in the higher group

According to Sudijono (1996:389), here is the classification and interpretation of discrimination index:

Table 3.3 Classifications and Interpretations of Discrimination Indices

Discrimination index	Classification
0.70-1.00	Excellent
0.40-0.70	Good
0.20-0.40	Satisfactory
≤ 0.20	Poor
Negative value on D	Very poor

The discriminating power of an item is reported as a decimal fraction; maximum positive discriminating power is indicated by an index of 1.00. This is obtained only when all students in the upper group answer correctly and no one in the lower group does. Zero discriminating power (.00) is obtained when an equal number of students in each group answer the item correctly. Negative discriminating power is obtained when more students in the lower group than in the upper group answer correctly. Both types of items should be removed and then discarded or improved.

5. The Effectiveness for Each Distractor

The researcher computed how well a distractor work by sticking on the computation of 5% of the total examinees number. Sudijono (1996:389) points out that a distactor can be said to have functioned well when it is chosen by the examinees at least 5% of the total number of examinees.

CHAPTER IV

FINDING AND DISCUSSION

This chapter presents findings of the research which include the validity, the reliability, the level of difficulty, the discrimination power, effectiveness of distractor and discussion.

A. The Description of Data

1. The Validity

The researcher used two types of validity, they were content validity and construct validity.

a) The Content validity

Firstly, the researcher analyzed the content validity of summative test items for the second grade students of MAN Tulungagung 1 in academic year 2013/2014. Content validity must be upon careful analysis of the language skill or an outline of the course and it is further expected the items to represent each proportion of the outline adequately. In addition, it was a comparison between what should be sampled by the test and what actually to be sampled. To know how good the content validity of summative test items for second grade students in MAN Tulungagung 1 was, the researcher compared the syllabus content to each test items as table 4.1:

Table 4.1 The Appropriateness of English Summative Test with The English Syllabus of MAN Tulungagung 1

Skills	The Materials in Syllabus	Item Number
Listening	1. Merespon makna dalam percakapan transaksional (*to get things done*) dan interpersonal (bersosialisasi) resmi dan berlanjut (*sustained*) secara akurat, lancar, dan berterima yang menggunakan ragam bahasa lisan dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: menyampaikan pendapat, meminta pendapat, menyatakan puas, dan menyatakan tidak puas	-
	2.Merespon makna dalam percakapan transaksional (*to get things done*) dan interpersonal (bersosialisasi) resmi dan berlanjut (*sustained*) secara akurat, lancar, dan berterima yang menggunakan ragam bahasa lisan dalam konteks kehidupan sehari-hari dan melibatkan tindak tutu: menasehati, memperingatkan, meluluskan permintaan, serta menyatakan perasaan *relief, pain,* dan *pleasure*	-
	3. Merespon makna yang terdapat dalam teks lisan fungsional pendek resmi dan tak resmi secara akurat, lancar dan berterima dalam berbagai konteks kehidupan sehari-hari	-
	4. Merespon makna dalam teks monolog yang menggunakan ragam bahasa lisan secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: *report, narrative,* dan *analytical exposition*	-
Speaking	1. Mengungkap-kan makna dalam percakapan transaksional (*to get things done*) dan interpersonal (bersosialisasi) resmi dan berlanjut (*sustained*) dengan menggunakan ragam bahasa lisan secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: menyampaikan pendapat, meminta pendapat, menyatakan puas, dan menyatakan tidak puas	5,6,9,11 and 13
	2. Mengungkap-kan makna dalam percakapan transaksional (*to get things done*) dan interpersonal (bersosialisasi) resmi dan berlanjut (*sustained*) dengan menggunakan ragam bahasa lisan secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: menasehati, memperingatkan, meluluskan permintaan, serta menyatakan perasaan *relief, pain,* dan *pleasure*	1,2,3,4,7,8,10,12,14 and 15
	3. Mengungkap-kan makna dalam teks lisan fungsional pendek resmi dan tak resmi secara akurat, lancar dan berterima dalam berbagai konteks kehidupan sehari-hari	-

Continued

Table 4.1
Continuation

	4. Mengungkap-kan makna dalam teks monolog dengan menggunakan ragam bahasa lisan secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: *report, narrative,* dan *analytical exposition*	-
Reading	1. Merespon makna dalam teks fungsional pendek (misalnya *banner, poster, pamphlet,* dll.) resmi dan tak resmi yang menggunakan ragam bahasa tulis secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari	16,17,18,19,20, 21,22,23,24 and 25.
	4. Merespon makna dan langkah retorika dalam esei yang menggunakan ragam bahasa tulis secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dan untuk mengakses ilmu pengetahuan dalam teks berbentuk: *report, narrative,* dan *analytical exposition*	26,27,28,29,30, 31,32,33,34,35, 36,37,38,39,40, 41,42,43,44 and 45
Writing	1. Mengungkap-kan makna dalam bentuk teks fungsional pendek (misalnya *banner, poster, pamphlet,* dll.) resmi dan tak resmi dengan menggunakan ragam bahasa tulis secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari	-
	2. Mengungkap-kan makna dan langkah retorika dalam esei dengan menggunakan ragam bahasa tulis secara akurat, lancar dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: *report, narrative,* dan *analytical exposition*	46.47,48,49 and 50

Based on the table above, it can be seen that not all the material in syllabus included in test items such as the third and fourth material in speaking, and the first material in writing. Moreover, not all material in listening included in test items.

From the table 4.1, it also can be taken the percentage of the skills being tested that represents the proportion of the content validity. Here is the table of the percentage of skills being tested:

Table 4.2 The Percentage of Skills Being Tested in Summative Test

The Language skills	The Percentage of Skills Being Tested
Listening	0/50 x 100 = 0 %
Reading	30/50 x 100 = 60%
Speaking	15/50 x 100 = 30%
Writing	5/50 x 100 = 10%

The table shows the skills of English test only represented reading, speaking, writing/grammar, while listening skill was not available. The table shows that the test items were dominated by 60% reading test. In other hand the speaking, writing and listening need to be practiced for achievement. Moreover, in syllabus material it covers the four skills, which must also be achieved by students.

b) The Construct Validity

The second analysis was construct validity. Hughes (1989:26) states that a test is said to have construct validity if it can be demonstrated that it measure just the ability which is supposed to measure. The words construct refers to any underlying ability which is hypothesized in a theory of language ability, so the researcher used the language testing theory to know whether the test has good construct validity or not. Here is the table presentation of techniques which were used in the test:

Table 4.3
The Techniques Used in English Summative Test

Speaking test The speaking test was shown in item numbers 1,2,3,4,5,6,7,8,9,10,11,12,13,14 and 15 • Item numbers 1,2,6,7,10,11 and 13 used the question about the meaning/ conclusion of the certain dialogue • Item numbers 3,4,5,8,9,14 and 15 used the blank dialogue and response the dialogue/expressions
Reading test The reading test was shown in item numbers 16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44, and 45 • Item numbers 16,17,19,21,22,23,24,25, 36,37, 39 and 42 provided the students to choose the correct answer related to the information of banner, advertisements and poster • Item numbers 18 and 34 was about deciding the similar meaning of the certain word • Item numbers 20, 40, and 41 provided the students to answer the purpose of certain functional texts • Item numbers 26,27,28,29,30,31,32 and 33 provided the students to identify the narrative text including the characters, the sequence events, the place etc • Item numbers 19, 34 and 43 provided the students to guess the meaning of certain unfamiliar words from context • Item numbers 35 and 38 provided the students to identify the reference
Writing test The writing test was shown in number 46,47,48,49,50 • Item numbers 46 and 47 provided the students to arrange the jumbled sentence into good paragraph • Item number 48,49 and 50 provided the students to complete the blank paragraph with the vocabulary provided

The technique of overall English skill test was multiple-choice question. The researcher found that speaking test was dominated by the questions about choosing the options which was appropriate with the blank dialogue, responding to certain dialogue and the meaning of dialogue.

In testing reading, the researcher found that the test provided students to choose the options consisting of information, the purpose of certain functional text, identify the character, place and sequence of events in narrative text. Then, the students had to choose the right answer for the vocabulary by finding unfamiliar vocabulary that was same with the vocabulary provided. The last was testing writing, the researcher found the arrangement of jumbled sentence where the students had to choose the option about sentence arrangement to be a good paragraph. The close test was also found in writing test. The students had to choose appropriate vocabulary which was appropriate with the paragraph.

2. The Reliability

The next was the reliability analysis. Reliability refers to the stability of the score. The reliability can be estimated by formula Kuder Richardson:

$\mathbf{r_{11}} = \left(\frac{n}{n-1}\right)\left(\frac{s_t^2 - p_1 q_1}{s_t^2}\right)$

$\mathbf{r_{11}}$ = reliability coefficient

n = number of test items

S_t^2 = standard deviation

p_1 = the right responds

q_1 = the wrong respond

Before computing the reliability, the standard deviation computed firstly as table 4.4:

Table 4.4 The preparatory to compute the standard deviation

No	Name	X_t	X_t^2
1	YK	41	1681
2	RF	41	1681
3	SN	40	1600
4	SND	40	1600
5	PM	34	1156
6	RN	39	1521
7	WP	40	1600
8	TA	40	1600
9	NA	38	1444
10	RM	38	1444
11	YNL	37	1369
12	SK	37	1369
13	WD	39	1521
14	LK	37	1369
15	FF	42	1764
16	FS	42	1764
17	ES	40	1600
18	DA	41	1681
19	AD	42	1764
20	AFS	42	1764
21	ML	41	1681
22	MKS	44	1936
23	GAR	41	1681
24	AF	39	1521
25	KM	41	1681
26	MAR	40	1600
27	LQ	41	1681
28	DTA	40	1600
29	AFA	43	1849
30	AW	40	1600
31	SA	37	1369
32	AM	36	1296
33	AZ	34	1156
34	EDS	33	1089
35	YS	37	1369
36	YN	25	625
37	YY	36	1296

Continued

Table 4.4
Continuation

38	YRS	36	1296
39	TM	37	1369
40	SA	37	1369
41	SR	37	1369
42	SNK	40	1600
43	SM	38	1444
44	SS	40	1600
45	RN	33	1089
46	RF	38	1444
47	RSA	38	1444
48	RA	33	1089
49	RS	31	961
50	RN	28	784
51	NI	40	1600
52	NM	39	1521
53	MH	28	784
54	MA	29	841
55	MAS	30	900
56	AK	29	841
57	AD	30	900
58	EHS	35	1225
59	MN	38	1444
60	KA	39	1521
61	IM	38	1444
62	IK	35	1225
63	IP	32	1024
64	BU	29	841
65	FH	34	1156
66	DW	35	1225
67	DS	30	900
68	HR	33	1089
69	MN	36	1296
70	LF	31	961
		$\sum X_t = 2574$	$\sum X_t^2 = 95918$

$$\mathbf{S_t^2} = \frac{\sum \mathbf{x_t^2}}{\mathbf{N}}$$

To know $\sum \boldsymbol{X_t^2}$ the formula below was used:

$$\sum \boldsymbol{X_t^2} = \sum \boldsymbol{X_t^2} - \left(\frac{\sum \boldsymbol{X_t}}{\boldsymbol{N}}\right)^2$$

$$= 95918 - \left(\frac{2574}{70}\right)^2$$

$$= 95918 - 94649.65$$

$$= 1268.4$$

Therefore, the standard deviation is

$$S_t^2 = \frac{1268.4}{70} = 18.12$$

After finding the result of standard deviation, the reliability can be computed by using Kuder Richardson formula (KR-20)

Table 4.5 The Table to Compote The Reliability by Using Kinder Richardson Formula (KR-20)

Item	Np	P_1	Nq	Q_1	P_1Q_1
1	13	0.185714	57	0.814286	0.1512243
2	56	0.8	14	0.2	0.16
3	1	0.014286	69	0.985714	0.0140819
4	37	0.528571	33	0.471429	0.2491837
5	9	0.128571	61	0.871429	0.1120405
6	3	0.042857	67	0.957143	0.0410203
7	8	0.114286	62	0.885714	0.1012247
8	66	0.942857	4	0.057143	0.0538777
9	64	0.914286	6	0.085714	0.0783671
10	10	0.142857	60	0.857143	0.1224489
11	50	0.714286	20	0.285714	0.2040815
12	43	0.614286	27	0.385714	0.2369387
13	20	0.285714	50	0.714286	0.2040815
14	41	0.585714	29	0.414286	0.2426531

Continued

Table 4.5
Continuation

15	52	0.742857	18	0.257143	0.1910205
16	6	0.085714	64	0.914286	0.0783671
17	0	0	70	1	0
18	8	0.114286	62	0.885714	0.1012247
19	6	0.085714	64	0.914286	0.0783671
20	9	0.128571	61	0.871429	0.1120405
21	3	0.042857	67	0.957143	0.0410203
22	19	0.271429	51	0.728571	0.1977553
23	8	0.114286	62	0.885714	0.1012247
24	11	0.157143	59	0.842857	0.1324491
25	2	0.028571	68	0.971429	0.0277547
26	2	0.028571	68	0.971429	0.0277547
27	1	0.014286	69	0.985714	0.0140819
28	6	0.085714	64	0.914286	0.0783671
29	14	0.2	56	0.8	0.16
30	7	0.1	63	0.9	0.09
31	61	0.871429	9	0.128571	0.1120405
32	4	0.057143	66	0.942857	0.0538777
33	25	0.357143	45	0.642857	0.2295919
34	7	0.1	63	0.914286	0.0783671
35	6	0.085714	64	0.914286	0.0783671
36	10	0.142857	60	0.857143	0.1224489
37	17	0.242857	53	0.757143	0.1838775
38	22	0.314286	48	0.685714	0.2155103
39	35	0.5	35	0.5	0.25
40	24	0.342857	46	0.657143	0.2253061
41	49	0.7	21	0.3	0.21
42	3	0.042857	67	0.957143	0.0410203
43	2	0.028571	68	0.971429	0.0277547
44	11	0.157143	59	0.842857	0.1324491
45	5	0.071429	65	0.928571	0.0663269
46	21	0.3	49	0.7	0.21
47	5	0.071429	65	0.928571	0.0663269
48	13	0.185714	57	0.814286	0.1512243
49	6	0.085714	64	0.914286	0.0783671
50	18	0.257143	52	0.742857	0.1910205
					$\sum p_1Q_1$=6.138161

Then, those scores above applied into kuder Richardson formula (KR-20)

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{S_t^2 - p_1 q_1}{S_t^2}\right)$$

$$r_{11} = \left(\frac{50}{50-1}\right)\left(\frac{18.12-6.138}{18.12}\right)$$

$$r_{11} = \left(\frac{50}{49}\right)\left(\frac{11.982}{18.12}\right)$$

$$r_{11} = (1.02)\ (0.66)$$

r11 = 0.6732

The result shows the reliability coefficient is 0.6732 0r 67%, it means that the reliability test is fair.

3. The Level of Difficulty

The level of difficulty shows how easy or difficult of test items. It can be seen through the number of students who can answer the items correctly. The level of difficulty can be estimated by using formula:

$$P = \frac{NP}{P}$$

Where,

P= level of difficulty

NP= the right response

N= the number of student

The level of difficulty index and classification can be estimated in the table 4.6 :

Table 4.6 The Presentation of Level Difficulty

Item	NP	N	P=NP/N	Classification
1	57	70	0.81	Easy
2	14	70	0.20	Difficult
3	69	70	0.98	Easy
4	33	70	0.47	Fair
5	61	70	0.87	Easy
6	67	70	0.96	Easy
7	62	70	0.88	Easy
8	4	70	0.06	Difficult
9	6	70	0.08	Difficult
10	60	70	0.86	Easy
11	20	70	0,28	Difficult
12	27	70	0,38	Fair
13	50	70	0.71	Easy
14	29	70	0.14	Difficult
15	18	70	0.25	Difficult
16	64	70	0.91	Easy
17	70	70	1	Easy
18	62	70	0.88	Easy
19	64	70	0.91	Easy
20	61	70	0.87	Easy
21	67	70	0.96	Easy
22	51	70	0.73	Easy
23	62	70	0.88	Easy
24	59	70	0.84	Easy
25	68	70	0.97	Easy
26	68	70	0.97	Easy
27	69	70	0.98	Easy
28	64	70	0.91	Easy
29	56	70	0.80	Easy
30	63	70	0.90	Easy
31	9	70	0.13	Difficult
32	66	70	0.94	Easy
33	45	70	0.64	Fair
34	63	70	0.90	Easy
35	64	70	0.91	Easy
36	60	70	0.86	Easy
37	53	70	0.76	Easy
38	48	70	0.69	Fair
39	35	70	0.50	Fair
40	46	70	0.65	Fair
41	21	70	0.30	Difficult
42	67	70	0.96	Easy

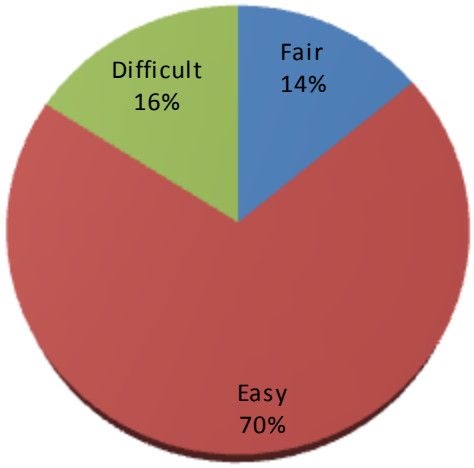
Continued

Table 4.6
Continuation

43	68	70	0.97	Easy
44	59	70	0.84	Easy
45	65	70	0.92	Easy
46	49	70	0.70	Fair
47	65	70	0.92	Easy
48	57	70	0.81	Easy
49	64	70	0.91	Easy
50	52	70	0.74	Easy

Based on the table 4.6, the percentage of the level of difficulty can be found as the following pie chart:

Figure 4.7 The Figure of The Level of Difficulty Percantage (English Summative Test Items of MAN Tulungagung 1)



4. The Discrimination Power

Discrimination power shows how well a test identifies differences in achievement level of students. The discrimination power of test items can be estimated by using formula:

D= PA-PB

Where,

D= Discrimination power

PA= Proportion of higher group

PB= Proportion of lower group

JA= the number of higher group

JB= The number of lower group

The discrimination power can be analyzed by classifying the students into three groups; upper group, middle group and lower (for detailed group position, see appendix IV). The researcher took only 27% of the lower and 27% of the upper group for this analysis and the rests belong to the middle group which was not taken to this analysis.

The researcher used discrimination index formula to find discrimination power criteria as table 4.8:

Table 4.8

The Data Presentation of Discrimination Power

Item	BA	BB	JA	JB	PA=BA/JA	PB=BB/JB	D=PA-PB	Classification
1.	19	13	19	19	1	0.68	0.32	Satisfactory
2.	5	6	19	19	0.26	0.31	-0.05	Very poor
3.	19	18	19	19	1	0.94	0,06	Poor
4.	15	5	19	19	0.78	0.26	0.52	Good
5.	19	13	19	19	1	0.68	0.32	Satisfactory
6.	19	16	19	19	1	0.84	0.16	Poor
7.	19	13	19	19	1	0. 68	0.32	Satisfactory
8.	0	2	19	19	0	0.10	-0.1	Very poor
9.	0	3	19	19	0	0.16	-0.16	Very poor
10.	19	11	19	19	1	0.58	0.42	Good
11.	6	6	19	19	0.31	0.31	0	Poor
12.	15	3	19	19	0.78	0.16	0.62	Good
13.	18	11	19	19	0.95	0.57	0.38	Satisfactory
14.	2	9	19	19	0.10	0.47	-0.37	Very poor
15.	0	2	19	19	0	0.10	-0.9	Very poor
16.	18	15	19	19	0.95	0.78	0.20	Satisfactory
17.	19	19	19	19	1	1	0	Poor
18.	19	14	19	19	1	0.73	0.27	Satisfactory
19.	18	15	19	19	0.95	0.78	0.17	Poor
20.	19	14	19	19	1	0.73	0.27	Satisfactory
21.	19	17	19	19	1	0.89	0.11	Poor
22.	12	12	19	19	0.63	0.63	0	Poor
23.	19	12	19	19	1	0.63	0.37	Satisfactory
24.	19	11	19	19	1	0.58	0.42	Good
25.	19	18	19	19	1	0.95	0.05	Poor
26.	19	18	19	19	1	0.95	0.05	Poor
27.	19	18	19	19	1	0.95	0.05	Poor
28.	18	17	19	19	0.95	0.89	0.06	Poor
29.	14	14	19	19	0.73	0.73	0	Poor
30.	19	17	19	19	1	0.89	0.11	Poor
31.	3	2	19	19	0.16	0.10	0.06	Poor
32.	19	17	19	19	1	0.89	0.11	Poor
33.	18	5	19	19	0.95	0.26	0.69	Good
34.	19	14	19	19	1	0.73	0.27	Satisfactory
35.	19	14	19	19	1	0.73	0.27	Satisfactory
36.	19	12	19	19	1	0.63	0.37	Satisfactory
37.	19	7	19	19	1	0.37	0.63	Good
38.	17	5	19	19	0.89	0.26	0.63	Good
39.	15	4	19	19	0.78	0.21	0.57	Good
40.	19	2	19	19	1	0.10	0.90	Excellent
41.	10	5	19	19	0.53	0.26	0.27	Satisfactory
42.	19	17	19	19	1	0.89	0.11	Poor
43.	19	17	19	19	1	0.89	0.11	Poor

Continued

Table 4.8
Continuation

44.	15	17	19	19	0.78	0.89	-0.11	Very poor
45.	19	15	19	19	1	0.78	0.22	Satisfactory
46.	17	12	19	19	0.89	0.63	0.26	Satisfactory
47.	19	15	19	19	1	0.78	0.22	Satisfactory
48.	16	13	19	19	0.84	0.68	0.16	Poor
49.	19	14	19	19	1	0.73	0.27	Satisfactory
50.	17	15	19	19	0.89	0.78	0.11	Poor

From the table above, the discrimination power for each item can be shown as the following pie chart:

Figure 4.9 The Percantage of Discrimination Power (English Summative Test Items of MAN Tulungagung 1)

5. The Effectiveness for Each Distractor

The effectiveness of distractor can be analyzed by finding out the number of students that choose the answers which they believed to be corrects but it was actually wrong answer. A distractor can be said to be well functioned if it has a strong power of attracting that it is chosen by at least 5% of the examinees. Here is the table of the effectiveness of distractor for each item. The symbol * represents the key answer, + represents the effectiveness of distractor, - represent

the distractors which are not effective and O stands for which no one chosen those and it must be revised.

Table 4.10

The Effectiveness of Distractor for Each Item

Item Number	Options	H (19)	M (38)	L (19)	H+M+L (70)	Percentage	Explanation
1	A	19	25	13	57	81%	*
	B	-	7	5	12	17%	+
	C	-	-	-	-	-	O
	D	-	-	1	1	1%	-
	E	-	-	-	-	-	O
2	A	5	3	6	14	20%	*
	B	-	1	1	2	3%	-
	C	14	28	12	54	77%	+
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
3	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	1	1	1%	-
	E	19	32	18	69	98%	*
4	A	-	-	-	-	-	O
	B	4	3	5	12	17%	+
	C	15	13	5	33	47%	*
	D	-	-	-	-	-	O
	E		16	9	25	36%	+
5	A	-	2	5	7	10%	+
	B	19	29	13	61	87%	*
	C	-	1	-	1	1%	-
	D	-	-	1	1	1%	*
	E	-	-	-	-	-	O
6	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	3	3	4%	-
	D	19	32	16	67	96%	*
	E	-	-	-	-	-	O
7	A	-	-	1	1	1%	-
	B	-	-	1	1	1%	-
	C	19	30	13	62	89%	*
	D	-	2	4	6	9%	+
	E	-	-	-	-	-	O

Continued

Table 4.10
Continuation

8	A	-	2	2	4	6%	*
	B	-	-	-	-	-	O
	C	19	29	17	65	93%	+
	D	-	-	-	-	-	O
	E	-	1	1	2	3%	-
9	A	19	25	15	59	84%	+
	B	-	4	1	5	7%	+
	C	-	3	3	6	9%	*
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
10	A		1	2	3	4%	-
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	19	30	11	60	86%	*
	E	-	1	6	7	10%	+
11	A	-	1	3	4	6%	+
	B	13	22	9	44	63%	+
	C	-	1	1	2	3%	-
	D	6	8	6	20	29%	*
	E	-	-	-	-	-	O
12	A	2	2	-	4	6%	+
	B	15	9	3	27	39%	*
	C	2	14	5	21	30%	+
	D	-	4	8	12	17%	+
	E	-	3	3	6	9%	+
13	A	1	9	1	11	16%	+
	B	-	-	4	4	6%	+
	C	18	21	11	50	71%	*
	D	-	-	-	-	-	O
	E	-	2	3	5	7%	+
14	A	14	4	7	25	36%	+
	B	-	-	-	-	-	O
	C	3	10	2	15	21%	+
	D	-	-	1	1	1%	-
	E	2	18	9	29	41%	*
15	A	6	4	2	12	17%	+
	B	-	3	8	11	16%	+
	C	13	7	6	26	37%	+
	D	-	2	1	3	4%	-
	E	-	16	2	18	26%	*
16	A	-	-	-	-	-	O
	B	-	-	4	4	6%	+
	C	1	1		2	3%	-
	D	-	-	-	-	-	O
	E	18	31	15	64	91%	*
17	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O

Continued

Table 4.10
Continuation

	D	19	32	19	70	-	*
	E	-	-	-	-	-	O
18	A	19	29	14	62	89%	*
	B	-	3	3	6	9%	+
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	2	2	3%	-
19	A	-	-	-	-	-	O
	B	-	1	3	4	6%	+
	C	1	-	-	1	1%	-
	D	-	-	1	1	1%	-
	E	18	31	15	64	91%	*
20	A	-	3	-	3	4%	-
	B	19	28	14	61	87%	*
	C	-	-	1	1	1%	-
	D	-	-	-	-	-	O
	E	-	1	4	5	7%	+
21	A	-	-	-	-	-	O
	B	-	-	2	2	3%	-
	C	-	-	-	-	-	O
	D	19	31	17	67	96%	*
	E	-	1	-	1	1%	-
22	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	7	5	6	18	26%	+
	D	-	-	1	1	1%	-
	E	12	27	12	51	73%	*
23	A	19	31	12	62	86%	*
	B	-	1	6	7	10%	+
	C	-	-	1	1	1%	-
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
24	A	-	-	5	5	7%	+
	B	-	-	-	-	-	O
	C	19	29	11	59	84%	*
	D	-	1	1	2	3%	-
	E	-	2	2	4	6%	+
25	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	1	1	2	3%	-
	D	19	31	18	68	97%	*
	E	-	-	-	-	-	O
26	A	19	31	18	68	97%	*
	B	-	-	1	1	1%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	1	-	1	1%	-
27	A	-	-	1	1	1%	-

Continued

Table 4.10
Continuation

	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	19	32	18	69	99%	*
28	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	18	29	17	64	91%	*
	E	1	3	2	6	9%	+
29	A	14	28	14	56	80%	*
	B	5	4	3	12	17%	+
	C	-	-	-	-	-	O
	D	-	-	2	2	3%	-
	E	-	-	-	-	-	O
30	A	-	-	-	-	-	O
	B	-	2	1	3	4%	-
	C	19	27	17	63	90%	*
	D	-	-	-	-	-	O
	E	-	3	1	4	6%	+
31	A	16	24	9	49	70%	+
	B	3	4	2	9	13%	*
	C	-	1	7	8	11%	+
	D	-	3	1	4	6%	+
	E	-	-	-	-	-	O
32	A	-	-	-	-	-	O
	B	-	-	1	1	1%	-
	C	-	2	-	2	3%	-
	D	-	-	1	1	1%	-
	E	19	30	17	66	94%	*
33	A	-	1	-	1	1%	-
	B	1	-	9	10	14%	+
	C	-	9	5	14	20%	+
	D	-	-	-	-	-	O
	E	18	22	5	45	64%	*
34	A	19	30	14	63	90%	*
	B	-	1	2	3	4%	-
	C	-	1	2	3	4%	-
	D	-	-	-	-	-	O
	E	-	-	1	1	1%	-
35	A	-	-	-	-	-	O
	B	-	1	4	5	7%	-
	C	-	-	-	-	-	O
	D	-	-	1	1	1%	-
	E	19	31	14	64	91%	*
36	A	19	29	12	60	86%	*
	B	-	-	-	-	-	O
	C	-	3	6	9	13%	+
	D	-	-	-	-	-	O

Continued

Table 4.10
Continuation

	E	-	-	1	1	1%	-
37	A	-	5	10	15	21%	+
	B	-	-	-	-	-	O
	C	-	-	2	2	3%	-
	D	19	27	7	53	76%	*
	E	-	-	-	-	-	O
38	A	2	6	10	18	26%	+
	B	-	1	2	3	4%	-
	C	-	-	1	1	1%	-
	D	17	25	6	48	66%	*
	E	-	-	-	-	-	O
39	A	2	-	-	2	3%	-
	B	15	10	4	29	41%	*
	C	-	2	2	4	6%	+
	D	2	20	13	35	50%	+
	E	-	-	-	-	-	O
40	A	-	1	4	5	-	-
	B	19	25	2	46	66%	*
	C	-	6	9	15	21%	+
	D	-	-	4	4	6%	+
	E	-	-	-	-	-	O
41	A	9	19	5	33	47%	+
	B	-	4	-	4	6%	+
	C	10	6	5	21	30%	*
	D	-	-	-	-	-	O
	E	-	3	9	12	17%	+
42	A	19	31	17	67	96%	*
	B	-	-	1	1	1%	-
	C	-	1	-	1	1%	-
	D	-	-	1	1	1%	-
	E	-	-	-	-	-	O
43	A	-	-	1	1	1%	-
	B	19	32	17	68	97%	*
	C	-	-	1	1	1%	-
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
44	A	-	-	-	-	-	O
	B	15	27	17	59	84%	*
	C		1	1	2	3%	-
	D	-	-	-	-	-	O
	E	4	4	1	9	13%	+
45	A	19	31	15	65	93%	*
	B	-	1	2	3	4%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	2	2	3%	-
46	A	-	-	-	-	-	O
	B	17	20	12	49	70%	*

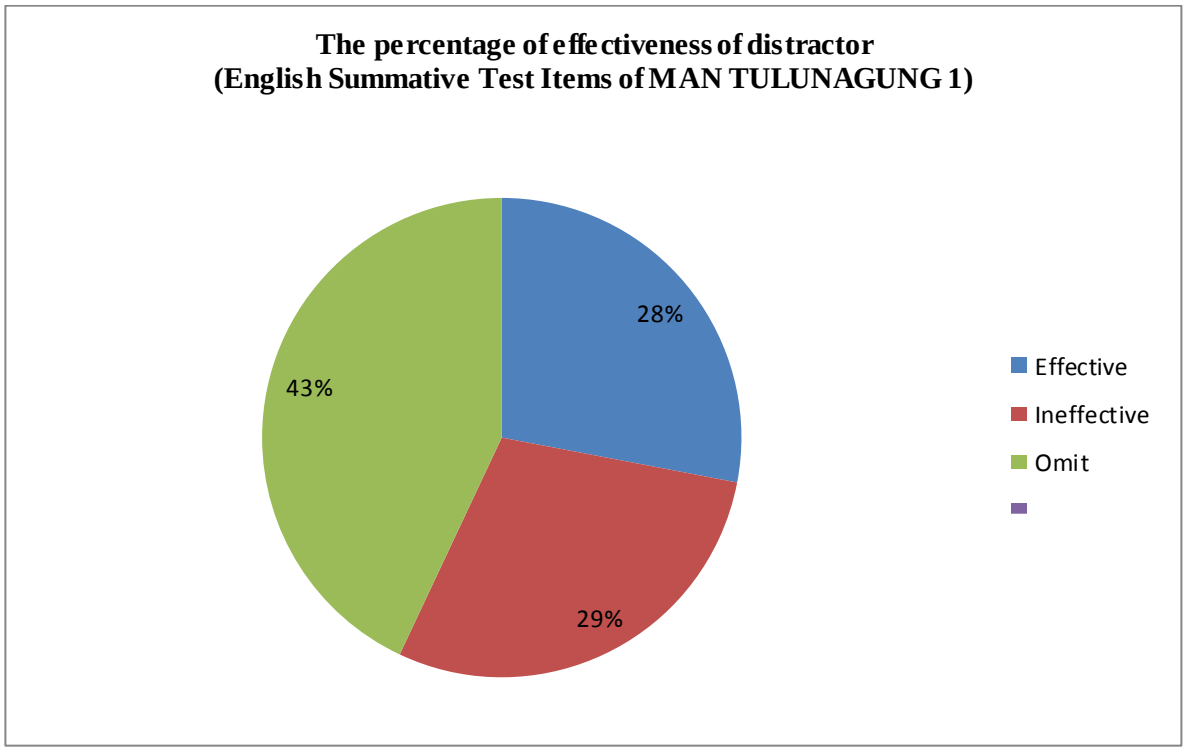
Continued

Table 4.10
Continuation

	C	2	12	7	21	30%	+
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
47	A	-	-	-	-	-	O
	B	-	-	1	1	1%	-
	C	19	31	15	65	93%	*
	D	-	-	-	-	-	O
	E	-	1	3	4	6%	+
48	A	-	-	1	1	1%	-
	B	16	28	13	57	81%	*
	C	2	-	-	2	3%	-
	D	-	-	-	-	-	O
	E	1	4	5	10	13%	+
49	A	19	31	14	64	91%	*
	B	-	-	4	4	6%	+
	C	-	1	-	1	1%	-
	D	-	-	1	1	1%	-
	E	-	-	-	-	-	O
50	A	-	-	1	1	1%	-
	B	-	2	-	2	3%	-
	C	17	20	15	52	74%	*
	D	2	10	3	15	21%	+
	E	-	-	-	-	-	O

Based on the table 4.10, some effective distractor was shown in option A in item numbers 5, 9, 11, 12, 13,14, 15, 24, 31, 37, 38 and 41. Option B in item numbers 1, 4, 9, 11, 13, 15, 16, 19, 23, 29, 33, 41 and 49. Option C in item numbers 2, 7, 8, 12, 14, 15, 18, 22, 31, 33, 36, 39, 40 and 46. Option D in item numbers 6, 12, 31, 39, 40 and 50. Option E in item number 10, 12, 20, 28, 30, 41, 44, 47, 48.

For more explanation, here is the data percentage of effectiveness of distractors :



2. Discussion

1. Validity

a. The Content Validity

The English summative test for second grade students of MAN Tulungagung 1 is not representative enough. It can be proved by table 4.1, not all the material for each skills included in the test. Whereas, a test will be a good content validity if the test contains materials taught to the students. Henning (2001:94) states " Content validity is concerned with whether or not the content of test is sufficiently *representative* and *comprehensive* for the test to be a valid measure of what it is supposed to measure".

From the table 4.2, it can also be shown the skill percentages as the representation of content validity are 0% of total items for testing listening, 60% of total items for testing reading, 30% of total items for testing speaking and 10% of total items for testing writing. It leads to be lack of content validity because there are four skills which have to be improved or achieved by the students, but the real test is only testing three skills. Indeed, the proportion is not fair. There are too many reading skills in the test.

The test maker should have attempt to quantify, balance the test components and assign a certain value to indicate the importance of each component in relation to the other components in the test. Heaton (1988: 161) states, "The test should achieve content validity and reflect the

components skill and area which the test maker wishes to include in the assessment"

b. The Construct Validity

The technique used in English summative test of MAN Tulungagung 1 was multiple-choice which assessed the three skills including speaking, writing and reading. The reading test used multiple choices as the technique of testing is appropriate enough with the language theory. But, the test was only purposed to test micro skills in reading like identifying referents of pronouns, using context to guess meaning and unfamiliar words and understanding relation among parts of text by recognizing indicators in discourse, especially for the introduction, development, transition and conclusion of ideas. Whereas, the reading test has two skills that are suggested to be tested, those are micro skill and macro skill.

The multiple-choice technique used in speaking and writing test, is not suitable enough with the theory of the language testing. The first is about speaking test; the student only chose the one of five options related to which one was suitable with the dialogue whereas the speaking proficiency usually deals with accent, grammar, vocabulary, fluency and comprehension. It is impossible to know the accent and fluency by only having multiple choices. The second is about writing test, according to the theory of writing test there are any sub- abilities such as control of punctuation, sensitivity to demand on style and so on. So that the sub

abilities of writing should have been considered more in writing test. The teacher will not absolutely be able to detect those sub-abilities by having multiple-choice technique.

A test, part of a test or testing technique is said to have construct validity if it can demonstrates that it measures only the ability which it is supposed to measure. Heaton (1988:161) states:

> "If a test has construct validity, it is capable of measuring certain specific characteristics in accordance with a theory of language behavior and learning, these types of validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills".

To make a good construct validity, it is supposed to use appropriate technique to assess the skills of language, do not always use the multiple-choice technique to assess all of the skills and components. Heaton (1988:161) states "…..if a communicative approach to language teaching and learning has been adopted throughout a course, a test comprising chiefly multiple choice items will lack construct validity".

2. The Reliability

The result of reliability coefficient of English Summative test for second grade of MAN Tulungagung 1 was 0.67. It is categorized as the fair reliability coefficients. Reliability is necessary characteristic of any good test: for it to be valid at all, a test must first be reliable as a measuring instrument. Reliability is thus a measure of accuracy, consistency,

dependability or fairness of scores resulting from administration of particular examination.

Many factors effect to the reliability. Brown (2004: 21-22) states that there are the factors affecting to the reliability; first, the most common reliability is caused by test-takers' temporary illness, fatigue, a bad day, anxiety and other physical or psychological factors. Second, it may be as the result of human error, subjectivity and bias. The next is caused by the condition which the test is administered. The last, the nature of test itself can cause measurement errors such as the length of test, the ambiguity options etc. Actually, the test can be purposed to be more reliable. Hughes (1988:36-43) suggests ways of achieving the more reliability test;

1. Take enough samples of behavior
2. Don't allow too much freedom
3. Don't write ambiguous items
4. Provide clear and explicit instructions
5. Ensure that tests are well laid out and perfectly readable
6. Test-taker should be familiar with format and testing techniques
7. Make comparisons between candidates as direct as possible
8. Provide a detailed scoring key
9. Identify candidates by number, not name
10. Employ multiple, independent scoring

3. The Level of Difficulty

The percentages of the level of difficulty from figure 4.7 were 70% easy items, 14% fair items and 16% difficult items. Item must be in appropriate difficulty for the students to whom it is administered. If possible, items should have indices of difficulty no less than 0.30 and no greater than 0.70. It is desirable to have most items in the 0.30-0.70 range of difficulty. Too difficult or too easy items contribute little to the discriminating power of a test. The level of difficulty for each item has the relationship and effect in arranging the test items.

From table 4.6, it can be seen that English summative for second grade students of MAN Tulungagung 1 has bad arrangement of difficulty level test. The test is started from easy question then followed by difficult question in number 2. Moreover, the easy item test appeared in the end item number.

Djiwandono (2008:220) states that giving the difficult question which makes the students think harder and consumes the more time to answer will lead to have bad effect, because they will feel inferior and afraid while doing the difficult items in the test and it also affect to the next questions. The difficult test items must be arranged in the last item so that the students feel confidence to answer because of having done the previous questions. Moreover, if the students feel troubled to answer the last items, it will not affect to the previous items.

4. The Discrimination Power

From the table 4.8 , it can be found that the test items for numbers 1, 4, 5, 7, 10, 12, 13, 16, 18, 20, 23,24, 33 34, 35, 36, 37, 38, 39, 40, 41, 45. 46, 47 and 49 included as the functioned discrimination, because they had the information about the differences in the performance of the students. The teacher or the test maker can keep saving those items to give in the next test. In other hand, the item numbers 3, 6, 11, 17, 19, 20, 21, 22, 25, 26, 27, 28, 29, 30, 31, 32,42,43,48 and 50 are considered poor discriminability. Because, the items cannot give the information about the differences of the performance among the students. Furthermore, the negative result in item number 2, 8, 9, 14, 15 and 44 are shown to have very poor discriminability. On contrary, the students who are supposed to have high ability got the wrong item. Besides, the students who are in lower ability got the correct item.

The discrimination is important feature of test. It is the capacity to discriminate among different candidates, reflect the differences in the performance of the individuals in the group and distinguish among the students who are in high ability or got the item correct and those who are in lower ability to respond the items correctly. The higher discrimination index of test item, the better it is.

Sudjiono (1996:408) state that following up must be done by the teacher or the test maker after analyzing the discrimination power for each item;

1. The items which have good discrimination power (satisfactory, good and excellent classification) should be kept in item test bank, so that can be used later
2. The items which are categorized as the poor discrimination should have been revised then used later or dropped
3. The very poor discrimination of test must be dropped or not to be used later

5. The effectiveness of Distractor for Each Items

Commonly, the multiple-choice question has the basic structure; a stem, option which consists of the answer and distractor. All of the incorrect options, or distractors, should actually be distracting. Preferably, each distractor should be selected by a greater proportion of the lower group than that of the upper group.

From table 4.10, the distractor A shows in item numbers 5, 9, 11, 12, 13,14, 15, 24, 31, 37, 38 and 41. Distractor B shows in item numbers 1, 4, 9, 11, 13, 15,16, 19, 23, 29, 33, 41 and 49. Distractor C shows in item numbers 2, 7, 8, 12, 14, 15, 18, 22, 31, 33, 36, 39, 40 and 46. Distractor D shows in item numbers 6, 12, 31, 39, 40 and 50. E shows in item numbers 10, 12, 20, 28, 30, 41, 44, 47 and 48. They are categorized as the effective

distractor, because there are more than 5% students who choose those distractors.

Besides that, distractor A shows in item numbers 7, 10, 20, 27, 33, 39, 48 and 50. Distractor B shows in item numbers 2, 7, 20, 21, 26, 30, 32, 35, 38, 42, 45, 47 and 50. Distractor C shows in item numbers 5, 6, 11, 16, 19, 25, 34, 37, 38, 43, 44, 48 and 49. D in item numbers 3, 14, 15, 19,, 22, 24,29, 35, 42, 45 and 49 . Distractor E shows in item numbers 8, 21, 26, 34 and 36. They are called as ineffective distactors, because those distractors are chosen by less than 5% from all the students. The other distractors which have not been mentioned above are called as omit because no students are interested in choosing. They should be deleted or revised

Sudjiono (1996:409) states that the distractor functions well while it is chosen by at least 5% from the all students. If a distractor elicits very few or no responses, then it may not be functioning as a distractor and should be replaced with a more attractive option. In addition, some distractor may be too appealing and causing the items to be too difficult. Very often items which have been rejected as having inappropriate difficulty, discriminability or variability can be redeemed by the revision of one or two response options.

Sudjiono (1996:410-413) states that if there is no one chooses the provided distractor, it means that the distractor cannot functioned well. It must be dropped or revised. Haladyna (2004:99) suggests how to write good options (either distractor or key answer) as follows:

1. Develop as many effective options as the test maker can, but two or three may be sufficient
2. Vary the location of the right answer according to the number of options. Assign the position of the right answer randomly
3. Place option independent; choices should not be overlapping
4. Keep the options homogeneous in content and grammatical structure
5. Keep the length of options about the same
6. Make distracters plausible
7. Use typical errors of the students when writing distractor
8. Avoid option that give clues to the right answer

CHAPTER V

CONCLUSION AND SUGGESTION

A. Conclusion

After analyzing the obtained data, five conclusions are deduced as follows:

1. English summative test of second grade of MAN Tulungagung 1 in academic year 2013/2014 was lack of content validity because the test did not include all the material stated in syllabus, it also did not include four skills provided in syllabus completely. The skills tested percentage were 0% items for testing listening, 60% for testing reading, 30% for testing speaking and 10% for testing writing.

However, based on the construct validity this test had good construct validity in reading test, because testing reading in the form of multiple choice was appropriate with the language theory. In speaking test, the multiple choice technique lead to have some flaws in construct validity because the tester could not evaluate the element which deal with the speaking test; pronunciation, fluency and comprehensions. Then multiple choices as the test formed in writing also lead to have low construct validity because the tester/teacher could not evaluate the factors dealt with writing test; punctuation, spelling sensitivity to demands on style, and so on

2. The coefficient of reliability in English summative test at MAN Tulungagung 1 was 0.67; it meant that the reliability of test was categorized as fair reliability test.

3. The level of difficulty of English summative test for second grade students of MAN Tulungagung 1 was 70% easy items, 14% fair items and 16% difficult items. It was dominated by easy items. So that the test items were too easy for the students. It lead to have bad of level difficulty.

4. The discrimination power of English summative test for second grade students of MAN Tulungagung 1 were 2 % excellent test items, 32 % satisfactory test items, 16 % good test items, 38% poor test items, 12 % very poor test items. Both satisfactory and poor discrimination power items dominated this test. It meant that the discrimination power for each of items was balance.

5. The percentage of distractors for each items in English summative test of second grade of MAN Tulungagung 1 were 43 categorized as omit, 28% effective distractor, and 29% ineffective distractor.

On the basis of the conclusion above, it could be drawn a general conclusion that the quality of English summative test for second grade students of MAN Tulungagung 1in academic year 2013/2014 was not good in term of both construct and content validity, the level of difficulty and the effectiveness of the distractors. Those aspects of test need to be improved.

B. Suggestion

According to the conclusion above, the English summative test for second grade students of Man Tulungagung 1 in academic year 2013/2014 was categorized as not good test, so that the researcher suggests the teacher or tester to consider the factors which affect to the quality of test in order to set the betterment or improvement to the next test.

Based on the weaknesses found in this study, the future researchers are suggested to conduct better study in terms of methodology and analysis, such as national examination, university entrance test and others.

REFERENCES

Ahmann, J Stanley and Glock, Marvin D. 1967. *Evaluating Pupil Growth, Principle of Test and Measurement*. : Boston :Allyn and Bason

Arikunto, Suharsimi.1997. *Prosedur Penelitian Suatu Pendakatan Praktik"* Jakarta:Rineka Cipta

Arikunto, Suharsimi.2012. *Dasar-dasar Evaluasi Pendidikan"*.Jakarta:Bumi Aksara.

Ary, Donald et al. 2002. *Introduction to Research in Education.*New York : CBS College Publishing

Bahman, Lyle F.1990. *Fundamental Considerations in Language Testing.* Toronto :Oxford University Press

Bailey, Kathleen.1998 *Learning about Language Assessment*. London:Heile Publisher

Bodgan ,R.C & Bikken.1988. Qualitative research for Education: An introduction to theories and methods. Boston:Pearson

Brown, H Douglas. *Teaching by Principle, An Interactive Approach to Language Pedagoy*. 2001. White Plains: Longman

Djiwandono,M.S. 2008. *Tes Bahasa*. Jakarta: Indeks.

Fraenkel,Jack. 2005. *How to Design and Evaluate Research in Education".* New York : McGraw-Hill Companies

Haladyna,Thomas.2004. *Developing and Validating Multiple-choice Test Items.* London: Lawrence Erlbaum Associates Publisher

Harrison,Andrew. 1983. *A language Testing Handbook*.. London: Macmillan Press

Heaton, J.B. 1988. *Writing English Language Tests*. London*:* Longman

Henning, Grant.1987. *A Guide to Languagege tests* London:.Longman

Hughes, Arthur.1989. *Testing for Language Teachers.* New York: Cambridge University Press

Lado,Robert.1983 *Language Testing.*London : Longman

Nitko,Anthony. *Educational Test and Measurement, an Introduction.* New York: Harcourt B Race Jovanovich.1983

Purwanto, Ngalim.1991. *Prinsip dan Tehnik Evaluasi Pengajaran.* Bandung: Remaja Rosda Karya

Rea-Dicksin, Pauline and Kevin Germaine.1992.*Evaluation* New York:.Oxford University Press

Sudjiono,Anas. *Pengantar Evaluasi Pendidikan.* 1996. Jakarta: Raja Grafindo Persada

Tanzeh, ahmad.2011 . *Metodologi Penelitian Praktis*. Yogyakarta: Teras

Vallete,Rebbeca M.1997. *Modern Language Testing.* New York.:Harcourt Brace Jovanovich Publisher

Weir, J Cyril.1990. *Communicative Language Testing.* New York : Prentice Hall.