ITEM AND RELIABILITY ANALYSIS OF THE ENGLISH SECOND SEMESTER FINAL TEST FOR EIGHTH GRADE STUDENTS AT SMPN 2 SEMARANG IN THE ACADEMIC YEAR 2017/2018

A FINAL PROJECT

submitted in partial fulfillment of the requirements
for the degree of *Sarjana Pendidikan*
in English Language Education

by
Zakiyah
2201414151

ENGLISH DEPARTMENT
FACULTY OF LANGUAGES AND ARTS
UNIVERSITAS NEGERI SEMARANG
2019

APPROVAL

This final project entitled *Item and Reliability Analysis of the English Second Semester Final Test for Eighth Grade Students at SMPN 2 Semarang in The Academic Year 2017/2018* written by Zakiyah (NIM 2201414151) has been approved by a team of examiners on 20 March 2019

Board of Examiners

1. **Chairman**

   Dr. Sri Rejeki Urip, M. Hum.

   NIP 196202211989012001

2. **Secretary**

   Dr. Rudi Hartono, M.Pd.

   NIP 196909072002121001

3. **Examiner I**

   Novia Trisanti, S.Pd., M.Pd.

   NIP 197611062005012002

4. **Examiner II**

   Puji Astuti, S.Pd., M.Pd., Ph.D.

   NIP 197806252008122001

   Puji Astuti

5. **Examiner III/ Supervisor**

   Rohani, S. Pd, M. A.

   NIP 197903122003121002

Approved by

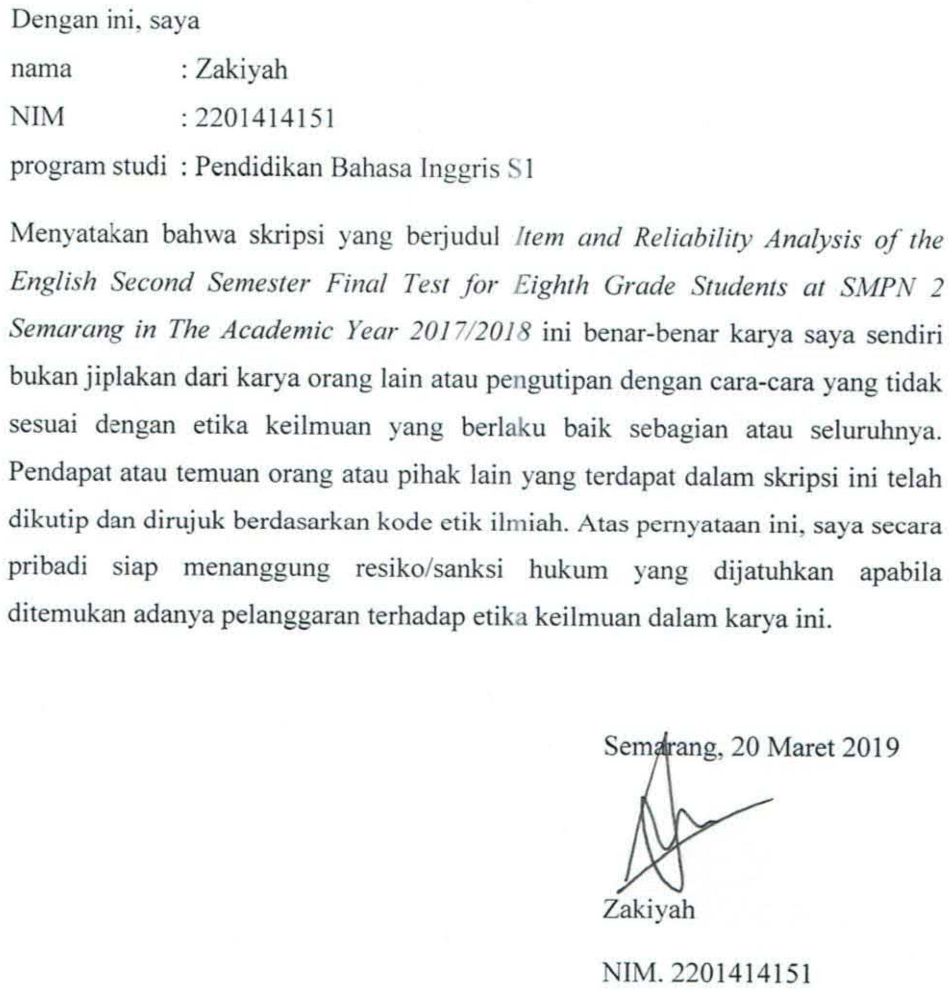
The Dean of Faculty of Language and Arts

Prof. Dr. M. Jazuli, M. Hum.

NIP 196107041988031003

KEMENTERIAN RISET, TEKNOLOGI DAN PENDIDIKAN TINGGI
UNIVERSITAS NEGERI SEMARANG
UNNES
FAKULTAS BAHASA DAN SENI

PERNYATAAN



Dengan ini, saya

nama : Zakiyah

NIM : 2201414151

program studi : Pendidikan Bahasa Inggris S1

Menyatakan bahwa skripsi yang berjudul *Item and Reliability Analysis of the English Second Semester Final Test for Eighth Grade Students at SMPN 2 Semarang in The Academic Year 2017/2018* ini benar-benar karya saya sendiri bukan jiplakan dari karya orang lain atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku baik sebagian atau seluruhnya. Pendapat atau temuan orang atau pihak lain yang terdapat dalam skripsi ini telah dikutip dan dirujuk berdasarkan kode etik ilmiah. Atas pernyataan ini, saya secara pribadi siap menanggung resiko/sanksi hukum yang dijatuhkan apabila ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya ini.

Semarang, 20 Maret 2019

Zakiyah

NIM. 2201414151

MOTTO AND DEDICATION

Do your best and let Allah do the rest

So be patient. Indeed, the promise of ALLAH is the truth (Q.S Ar-Rum: 60)

Allah does not burden a soul beyond that it can bear. (Al-Baqarah: 286)

Study while others are sleeping; work while others are loafing; prepare while others are playing; and dream while others are wishing. (William Arthus Ward)

I dedicate this final project to:

- My beloved mother and father
- My dearest siblings
- My lovely best friends

ACKNOWLEDGEMENT

Bismillahirrahmanirrahim, first and foremost the researcher would like to extend her gratitude to The Most Gracious and Merciful Allah SWT for the blessing, inspiration, and health leading to being able to finish the final project.

The researcher gave her deepest appreciation to the following people:

1. For Rohani, S.Pd, M.A. as her advisor, for the guidance, kindness, valuable advice, indispensable helpful correction and unfailing encouragement in the process of making and completing this final project.
2. For the head of the English Department, Dr. Rudi Hartono MP.d. and the head of the English education study program, Galuh Kirana Dwi Areni MP.d for their help in facilitating the students' academic.
3. For all lecturers of English Department who had taught and shared their brilliant knowledge and precious lesson during 4 years of her college.
4. For Mr. Siminto as Headmaster, Ma'am Setyo Asri and Ma'am Kusmawarni as English teachers, administrative staff and the eighth grade students from SMPN 2 Semarang for their help, contribution, for giving a permission and for providing the data needed during this research process conducted in their school.
5. For Her beloved family, her father, her mother, her sister and her brother for their pray, help, endless love, moral support, valuable guidance, and unfailing encouragement in finishing her study.
6. For Her best friends at Universitas Negeri Semarang for their endless love, the greatest spirit, hard effort, sincere support, and share unforgettable moments together in completing this study.

The researcher

ABSTRACT

Zakiyah. (2019). *Item and Reliability Analysis of English Second Semester Final Test for The Eighth Grade Students of SMPN 2 Semarang in the Academic Year 2017/2018.* Final Project, English Language Education, Faculty of Languages and Arts, Semarang State University. Advisor: Rohani, S.Pd, M.A.

Keywords: *item difficulty, item discrimination, alternatives , and reliability*

A test is used to measure students' achievement in a certain period. By testing, the teachers will know their students' understanding, or difficulty of their teaching learning process at that time, so they can improve their teaching method, media, material, and assessment. Beside that, a test must be a good measure tool to identify the students' achievement well. Therefore, the test had to be limited errors as small as possible by trying out to reveal the characteristics of multiple choice test.

The aim of this study was to reveal the characteristics of English second semester test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018. The characteristics of the multiple choice test are item difficulty, item discrimination, alternatives which are answer key and distractor, and reliability.

In conducting this study, descriptive analytic and quantitative method were used to reveal the characteristics of the multiple choice test. The data analysis technique used the Item and Test Analysis Program (ITEMAN) version 3.00. The subject of this study was the eighth grade students of SMPN 2 Semarang (287 students as population and 60 students as sample taken by random sampling). The data was obtained from an interview with the English teacher of SMPN 2 Semarang and the head of English MGMP *Sub Rayon* 01 East Semarang. Furthermore, the documents used in this study were English second semester test papers, answer key and students' answer sheets.

The result of this study showed that: (1) The item difficulty attained 23 items (57.5%) which were considered as easy category, 12 items (30%) as medium category, 5 items (12.5%) as difficult category. (2) The item discrimination attained 9 items (22.5%) which were considered as bad category, 9 items (22.5%) as sufficient category, 17 items (42.5%) as good category, 5 items (12.5%) as excellent category. (3) Based on the alternatives, the English second semester final test had 65 ineffective distractors and 55 effective distractors. The items which had all of the effective distractors was just 7 items. For answer keys, there were 2 answer keys of 40 answer keys which should be cross-checked. They were item number 3 and item number 38. (4) The reliability of the test was 0.717. By analyzing the characteristics of the multiple choice test, the teachers as test maker would produce a trusted test to measure the students' achievement accurately and improve their teaching learning method.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF APPENDICES

CHAPTER I
INTRODUCTION

Chapter one presents background of the study, reason for choosing the topic, research questions, objectives of the study, significance of the study, limitation of the study, definition of key terms, and the outline of the study.

1. 1 Background of the Study

English is a primary foreign language for Indonesian that must be mastered by the students since it is very essential in this globalization era to communicate with people around the world. In modern era, it has already been taught from playgroup until university. Every student is supposed to reach the next level by testing them to deal with new and more advanced to the learning objectives. For instance, the semester final test is one of the requirements that must be passed by students to gain the higher level. Without any test, the teachers will find difficulties to give evidence of the quality of their students to the next stage, therefore, they must conduct test to assert their students' knowledge or ability.

Jandaghi, G. (2011, p. 1) stated that testing is an important part for teachers to be able to evaluate their students during the end of an educational course in a teaching-learning process. By testing, teacher is able to get information about how far students understand the material of certain subject and improve the teaching learning process in the teaching method, media, material, and assessment. Shomami (2014, p. 1) stated that the objective of a test give about students' progress information for a teacher to ascertain how far the goals of learning have been achieved and reviewed the effectiveness of teaching method in the teaching learning process. The statement showed that it is very useful for the student in their learning and for the teacher in their teaching. As stated Roszkowski & Spreat (2011, p. 13), test is a systematic procedure to obtain information about a person, object, or situation. It can be used as feedback for teachers in improving and editing learning programs and learning activities, while for students, it can represent whole of their performance in learning teaching process. A score of the test shows their achievement.

The main purpose of testing in education attribute is to identify the real characteristics of an object such as students' ability, achievement, behaviour, passion, or unique measure indirectly. To measure the real score, it needs a good instrument, thus, they are able to identified well. A good test must consist of good items which fulfill requirements based on the characteristics of multiple choice test, and it must give a real information containing errors as small as possible (Mulianah & Hidayat, 2013; Suwarto, 2016). According to Surapranata (2004, p.88) and Nugiyantoro, Gunawan, & Marzuki. (2002, p. 320), every student will obtain score that consist of three part, observed test score, true score that is the students' real ability, and error of measurement, therefore, the error of measurement must be reduced to improve qualified test. For instance, summative test often uses multiple choice form and each test item must be good item. The test is expected to have error as little as possible, thus, the test can measure students' achievement accurately. It must be trusted to measure students' achievement, so the characteristics of multiple choice test must be analyzed accurately and Suwarto (2007, p. 168; 2016, p. 3) stated that the characteristics of multiple choice test item must have an adequate item difficulty, a good item discrimination, and distractor functioning. Moreover, reliability test is also important because the higher reliability index of the test, the test will be properly to measure students' achievement. According to Nugiyantoro et al. (2002, p. 320 & 334), if the reliability index of a test is high, the test will able to measure minimize the smallest possible error scores in the test to measure students' achievement properly. Then, if the reliability index of a test is high, the test items of the test will be good and responsibility to measure students' achievement. It is also supported by Surapranata (2004, p. 10). He stated that quantitative analysis of test item is the internal characteristics analysis of the multiple choice test through empirical data. The quantitative internal characteristics analysis is item difficulty, item discrimination, and reliability. In addition, for multiple choice form, there are students' answer that they guess or answer correctly and the effectiveness of distractors. Both of them are dissemination of alternative for the students tested. Therefore, forming multiple choice test must be considered the item difficulty, item discrimination, alternatives, and reliability. By analyzing them, the test can distinguish between students who have high achievement and low achievement Surapranata (2004).

Surapranata (2004, p. 11) stated that one of the purposes in analyzing test items is to improve the qualified test that are (1) it can be used because of proven to be good item by supporting numerical data analyzed statistically, (2) it should be revised because of weakness of the test item, (3) it must be eliminated because it has not function empirically. According to Masruroh (2014, p. 3), teachers analyzing test items will be able to know which one good item or bad item, they will be able to distinguish between low, middle, and high students' achievement. Therefore, by analyzing the test, the teacher can be able to determine which item that can be used and saved in bank test and which item that should be edited or dropped.

1. 2 **Reason for Choosing Topic**

In this study, the researcher had interviewed one of the English teachers of SMPN 2 Semarang that could be seen in appendix 1. Based on the interview, the mid-term test was designed by the teacher herself. However, the final test was made by the teacher team of a district or *Sub Rayon* called MGMP (*Musyawarah Guru Mata Pelajaran*). The team had a responsibility to design a test for each subject. One of the subjects was English. The researcher had interviewed the head of English MGMP *Sub Rayon* 01 East Semarang as well which could be seen in appendix 2. Based on the interview, the English second semester final test of the eighth grade students made by the English MGMP *Sub Rayon* 01 was not tried out. On the process of making the final test, they did not analyze the item difficulty, item discrimination, its alternatives (answer key and distractions) of each test item, even the reliability of the test. As stated by the head of English MGMP *Sub Rayon* 01, there was an editing process where the English MGMP *Sub Rayon* 01 cross checked the test items each other, then the test had already been distributed to students. It showed that the test was not trusted to measure the students' achievement because it was analyzed yet.

Therefore, the researcher was interested in analyzing the characteristics of English second semester final test's multiple choice for the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018. A test maker should try out the test before it was done by the test-takers so that the test could be believed as a measuring tool. To achieve the goals above, the researcher carried out the current research by "Item and Reliability Analysis of English Second Semester Final Test for the Eighth Grade Students of SMPN 2 Semarang in the Academic Year 2017/2018." This study was concerned with the English second semester final test for the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018 designed by the English MGMP *Sub Rayon* 01 East Semarang.

1. 3 **Research Questions**

Based on the background presented and the reason above, the researcher formulated four questions of this study below:

1. How is the item difficulty of each test item of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018?
2. How is the item discrimination of each test item of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018?

3. How are the alternatives (distractor and answer key) of each test item of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018?
4. How is reliability of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018?

1. 4 **Objectives of the Study**

The objectives of this study based on the research questions above were as follows:

1. To reveal the item difficulty of each test items of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018.
2. To reveal the item discrimination of each test items of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018.
3. To reveal the qualified alternatives of each test items of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018.
4. To reveal the reliability of English second semester final test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018.

1. 5 Significance of the Study

This study was expected to give some information to:

1. Teachers

   For teachers, they could learn this study as a reference for what the characteristics of multiple choice test are and how to analyze them. Analyzing test makes the test to be believed as a good measuring tool based on standard criteria, namely, item difficulty, item discrimination, alternatives, and reliability. If the test could be believed, it would be useful to improve their evaluation of their teaching learning process, diagnose the effectiveness of their teaching, and know the result of their students' achievement during certain period accurately.

2. Other Researchers

   For other researchers, they could use this study as a reference for their study related to what the characteristics of multiple choice test are and how to analyze them. It was also used to develop a test item analysis process on other subjects.

1. 6 Scope of the Study

The test item of English second semester test of the eighth grade students of SMPN 2 Semarang in the academic year 2017/2018 that was made by English MGMP *Sub Rayon* 01 East Semarang consists of 45 items in 40 multiple choices questions and 5 items in essay. However, the researcher just focused on the multiple choice form which certainly have only one correct answer. Thus, the answer will be objective for its correction. Moreover, the test item and the reliability of test would be easy to be computed and analyzed.

1. 7 Definitions of Key Terms

The key terms in this study were item analysis, item difficulty, item discrimination, alternatives, reliability, and test.

1. Item Analysis: a statistical method used to determine the qualified test based on the characteristics of the multiple choice test by analyzing each item test in three aspects; item difficulty, item discrimination, and alternatives.
2. Item difficulty: A comparison between the students answered correctly of a item and the total of students did the test.
3. Item discrimination: item ability to distinguish between the upper group and lower group.
4. Alternatives: the options of multiple choice test that consist of answer key as a true answer and distractors as false answer to confuse students when they choose.
5. Reliability: the qualified test of being trust or of performing consistently well.
6. Test: an instrument, a measuring tool, a method or procedure to obtain development and growth students' performance.

1. 8 **Outline of the Report**

The study was organized into five chapters consisting of an introduction of this study, review of related literature, research methodology, findings and discussions of the study, and the last chapter was conclusions and suggestions. The followings are detailed explanation of each chapter.

Chapter I covers the background of the study, the reason for choosing the topic, research questions, objectives of the study, significance of the study, limitation of the study, definition of key terms, and the outline of the study.

Chapter II reviews previous studies and explained the theoretical background, the figured framework of the present study in diagram form.

Chapter III describes the research design, the research site, the subject of the study, the object of the study, the research variables, the procedures of data collection, and the technique of data analysis.

Chapter IV reports the result of the research and discusses the result with other studies.

Chapter V provides some conclusions and suggestions for this study

CHAPTER II
REVIEW OF RELATED LITERATURE

Chapter two reviews previous studies related to this study; item difficulty, item discrimination, alternatives, and reliability analysis conducted in the past. It also explained theoretical background which contained definition and quotations about item, alternatives, and reliability analysis. Moreover, this chapter figured framework of the present study in diagram form which was summary of the theoretical study.

2. 1 Review of Previous Studies

There were a number of studies conducted in the past related to this study. These studies analyzed the characteristics of multiple choice test, namely, item difficulty, item discrimination, alternatives, and reliability.

Firstly, the researchers analyzed item difficulty (p). Singh, Kariwal, Gupta & Shrotriya (2014) analyzed the quality of multiple choice questions through item analysis. He found 11 (5%) items that belonged to medium category with range 30 % - 70 %, 9 (45%) items that belonged to easy category with range p> 70 %, and no any items that belonged to difficult category with range p< 30%. Chauhan (2013) calculated difficulty index of each individual item in stem type multiple choice question of anatomy subject. They found 35 items of 65 items that belonged to acceptable range (30-50% or 60-70%), 3 items of 65 items that belonged to difficult category (p<30%), 12 items of 65 items that belonged to easy category (p>70%), and 15 items of 65 items that belonged to ideal quality

(50-60%). Most of the items were acceptable difficulty index. Suruchi & Rana (2014) analyzed item difficulty of test items in an achievement test in Biology. They found 1 item of 120 items that belonged to difficult category with range below 0.20, 18 items of 120 items that belonged to good category with range 0.20-0.50, 94 items of 120 items that belonged to best category with range 0.50-0.80, and 7 items of 120 items that belonged to very easy category with range above 0.80. Thus, they determined that the one difficult item and seven easy items should rejected for the final draft of achievement test. Kolte (2015) found 4 difficult items with range p< 30%, 26 acceptable items with range 30-70%, and 10 easy with range p>70% items in item difficulty. Sa'adah (2017) analyzed the items quality of English Mid-Term Test. She found 18 items (72%) as ideal category with range around 0.62, 2 items (8%) as easy category with range p> 0.90 and 5 items (20%) as difficult category with range p< 0.20. Saputra (2015) compared the quality of the English second mid-term test between SMP N 1 Semarang and SMP Kesatrian 2 Semarang. He found 31 easy items, 15 moderate items, and 4 difficult items of SMP N 1 Semarang, while, in SMP Kesatrian 2 Semarang, there were 36 easy items and 14 moderate items. Those researchers above analyzed item difficulty through a certain formula such as difficulty index (p) or prop correction, it included that they analyzed it manually, however, there were the researchers who used a program to analyze it, for instance, Mulianah & Hidayat (2013) used ITEMAN version 3.00 program to analyze item test of computer based test including item difficulty. In addition, each researcher chose a certain theory of categories to determine the quality of items.

Secondly, the researchers analyzed item discrimination (D). It is one of item analysis which can be calculated its indexes manually or through computer software such as SPSS (Statistical Product and Service Solutions), Microsoft Excel, ANATES program, and ITEMAN (Item and Test Analysis Program). For instance, Chellamani & Boopathiraj (2013) had analyzed item discrimination with separating method between upper groups and lower groups which their scores entered in Microsoft Excel. Zubairi & Kassim (2006) used SPSS and BIGSTEPS to analyze the item characteristics that were the item measurring tool difficulty and the item discrimination indexes. Another example, Raharja (2014) analyzed item discrimination with ANATES V4. In her study, there was not the excellent category of item discrimination. There were just 8 items in the good category, for the sufficient category was 13 items, and the bad category was 28 items. Therefore, the bad items had to be dropped, and sufficient items should be revised. Moreover, bad item showed that the item could not distinguish between students had mastered the competencies and students who had not mastered competence. Every researcher had their own theory of their category and the recommendation of the categories from the existing theory depended on item test types.

Thirdly, the researchers analyzed distractor. Putri (2015) analyzed it with ITEMAN version 3.00 program, however, for reliability, validity, item difficulty, and item discrimination, she analyzed them manually. It indicated that she did not know that ITEMAN version 3.00 program could analyze all of them used ITEMAN version 3.00 program. As this had already been done by Rusmiana

(2015). She used ITEMAN version 3.00 program to determine the characteristics of a good test.

The last characteristic of a test is reliability. A test is said reliable if the test is consistent from time to time to produce the same score. Many studies conducted in the past who found the reliability index used scientist's formula that often used Kuder - Richardson 20/21 formula (KR-20/ KR-21). The researchers found a reliable test or not reliable test with certain formula were Bernasela (2014), Haryudin (2015), Pascual (2016), and Sugianto (2017). Pascual (2016) described that the English achievement test for ESL learners in Northern Philippines was reliable. Nevertheless, there were a researcher, Hidayati (2009) who found moderate reliability of the English mid-term test of eighth grade students of the SMP 33 Semarang.

The difference between those studies above and this study was technique data analysis which used ITEMAN version 3.00 program to analyze and reveal item difficulty, item discrimination, distractor even answer key. Nevertheless, it could also be similarity because Rusmiana (2015) also used ITEMAN version 3.00 program to analyze the characteristic of a test. The difference between Rusmiana's study and this study were the object of study. Rusmiana's study analyzed Accounting theory for vocational education, while, the object of this study was the summative test which was English second semester final test of the eighth grade students.

Another similarity between those studies above and this study was a descriptive quantitative approach. Almost of those studies used the descriptive quantitative method. For another difference was answer key analysis. From those studies above, there were rarely discussing answer key, so in this study would be analyzed answer key.

The recommendation of the past studies above to next other researcher was the easy method to analyze item analysis (item difficulty, item discrimination, and alternatives). There are some software that can be used for analyzing in modern area, such as, Software ITEMAN, Anates, Microsoft Excel, SPSS (Statistical Program for Social Science). For this study, the researcher chose ITEMAN version 3.00 program to analyze item analyze and its reliability.

2. 2 **Theoretical Background**

This theoretical background discussed some theories related to the topic of this study. There were two main points that would be discussed in this part, namely, test and the characteristics of multiple choice test.

2. 2. 1 A Test

Test is one of instruments to collect information about students' performance that consists of items that must be answered by students to obtain their information.

2. 2. 1. 1 The Definition of Test

Test is one of measuring tools to obtain information about educational attribute. According to Allen & Yen (1979, p. 1) and Brown (2004, p. 3), a test is a set of procedure to measure an individual's ability, knowledge or performance. Roszkowski and Spreat (2011, p. 13) also argued that a test is a systematic procedure to collect information of persons, objects, or situations and to score of the collecting data. A test has an important role in educational tools, for instance, the test helps teachers to evaluate and assess students' achievement what they have learned in the learning process. The teachers can also use test to motivate and help students' academic efforts (Jandaghi, 2011). By testing students, they indirectly are motivated to study hardly.

According to Brown (2004, p. 3), there are some components of a test. Firstly, the method of a test must be explicit and structured to qualify as a test. It is like multiple-choice questions with determining correct answer key, a writing prompt with scoring rubric, and an oral interview with question script. Secondly, a test must measure general ability, while others focus on very specific competencies. A multi-skill proficiency test defines a general ability level. A quiz of certain material measures specific knowledge. Thirdly, a test must measures test-taker's performance.

Based on the explanation above, the researcher concluded that the test is a measuring tool, an instrument, a method or procedure to obtain development and growth of students' performance such as in English language; speaking, writing, listening, reading, and subset of English language in a certain period. Test provides an accurate measure tool of student's ability or behavior.

2.2.1.2 Significance of Test

Harydin (2015, p. 79) stated that test has many reasons why students must be tested:

a. Achievement

Achievement is an action that can achieve a thing successfully by effort, courage, or skill. A test can be used to evaluate a student' achievement that must suit with the learning objectives.

b. Motivation

Motivation is an internal desires that refers to derive behavior to which pushes someone to do things in order to achieve goals and directs the individual activities. (Lai, 2011, p. 6). By taking test, students can motivate themselves to do it in order to get high score.

c. Encouraging students

By testing, students will get bad score, medium score, or high score. It makes them encouraged to study hard.

d. Diagnosis

Diagnosis can determine the strengths and weaknesses of students in learning (Hughes, 2003), for example, for low students teacher can give a remedial test.

e. Experimentation

Test can be used in educational experiments such as pre-test and a post-test that are given in experimental and control class to know the effectiveness of a certain technique of teaching.

f. Promotion and Advancement

Every grade from from lower class to upper class of formal school must certainly hold a test. It is called promotion and advancement of each student.

g. Parents Information

Students' parents will get report of students' achievement which contains test scores of midterm test and final semester. Therefore, the parents know their children's achievement.

2. 2. 1. 3 Purposes of Test

Generally, a test has many purposes, Suwarto (2004, p. 290) stated that there two categories of purpose of test which are bureaucratic and professional categories. Firstly, the bureaucratic category is to control, monitor and certify in attainment of summative functions. Secondly, the professional category related to student learning in which the teacher's ability to determine whether the development and understanding has been accepted by the student, whether the teaching learning process is effective or not. In addition, According to Djemari (2004, p. 72), the important purposes of a test:

a. Knowing the level of student ability
b. Measuring the students' growth and development
c. Diagnosing students' learning difficulties
d. Knowing the results of teaching learning program
e. Knowing the learning outcomes
f. Knowing the achievement of the curriculum
g. Encouraging students to learn
h. Motivating teacher to have better teaching method

2. 2. 1. 4 The Types of Test

Based on the purposes above, there are four types of test that are widely used in educational institutions. There are four types of a test; placement test, diagnostic test, formative test or achievement test, summative test or proficiency test (Brown, 2004; Suwarto, 2013):

a. Placement test are carried out at the beginning of a lesson. It can determine the level of student ability. Whether the student needs matriculation, additional lessons or not, it is determined from this test. In addition, Rudyatmi & Rusllowati (2017, p. 14) stated that the purpose of conducting this test is to place students in certain programs that match with their characteristics. It means that the test of beginning of the lesson measures student preparation and knows their knowledge which had been achieved of certain lesson. Moreover, placement test in English lesson has many varieties; assessing comprehension and production, responding through written and oral performance, open-ended and limited responses, selection (multiple choice) and gap-filling formats depending on the curriculum (Brown, 2004).

b. Diagnostic test can be used to know learning difficulties faced by students including misconceptions in teaching learning process. The test is held because most students fail in teaching learning process. Its result provides the information whether students have mastered the material or not. Thus, the teachers can know student learning difficulties, and teachers can improve the specific way to give treatment for their students. For instance, a pronunciation test might diagnose the phonological features of English which are difficult for students, so the material must become inside of curriculum (Brown, 2004).

c. Formative test or achievement test is to obtain feedback on the level of success of the implementation in learning process. The feedback is useful to improve teaching strategies. This test is conducted periodically throughout the semester. The material of the test is chosen based on each chapter or sub chapter. In addition, Rudyatmi & Rusllowati (2017, p. 13) stated the formative test is useful for students and teacher. For teacher, if the class failed in formative test result, it would be input to revise learning programs that have been prepared; methods, media, time allocation in teaching learning process. By revising, the learning program will be effective, appropriate with condition, ability, passion of students.

d. Summative test or proficiency test is given to students at the end of semester. The result determines student's learning success which is score and knows certain ability mastered by students. In addition, Rudyatmi & Rusllowati (2017, p. 12) stated that summative test is not only a test carried out at the end of each semester, but also at the end of each module, final school examinations, national exams, and even semester final exams in university.

In addition, Brown (2004, p. 43) sated that language aptitude test is a kind of tests that was designed to measure general ability to learn a foreign language and ultimate success in that undertaking. This test is belonging to test types above. designed to classroom learning of any language. However, the test type which was analyzed by the researcher was summative test, namely, second semester final test.

2. 2. 1. 5 The Item Test Types

Test can be divided to some item test types based on various point of views. Rudyatmi & Rusllowati (2017, p. 23-49) and Suwarto (2013, p. 34-58) divided test into two general categories, namely, objective test and subjective test (essay test). The objective test consists of items which have only one right answer, meanwhile, the subjective test requires long sentences answers such as explanation, compared, mention differences and similarities. This test can reveal how students remember, understand, and organize their ideas or a lesson that have been learned, by expressing them in their written form with their own words.

a. The advantages and disadvantages of objective tests and subjective test (essay test)

   1) There are some advantages of using objective tests:

      a) The way of scoring is easier, fast, and accurate, for instance, one true answer is one score. Therefore, everyone can score it. Whenever its scoring, the score is still same.

      b) The test can represent the whole of curriculum material, so the content validity can be more responsible and the test will represent students' knowledge of lesson tested.

      c) Students can answer quickly, so they may be able to answer all of the test item.

2) There are some disadvantages of using object tests:

   a) The test will be difficult to be made by teacher, because the teacher must make distractors to make confuse when students choose the correct one of the alternatives.

   b) This test requires expensive cost and long time to print the test paper.

   c) The students randomly suggest the true answer of alternatives. It means that students might not know which one the really true answer is.

3) There are some advantages of using subjective tests:

   a) The test can measure students' mental process.

   b) The test is easy to be made by teachers because the test item is limited.

4) There are some disadvantages of using subjective tests:

   a) This test cannot represent the whole curriculum material because test item is limited.

   b) The way of scoring is subjective. The subjective scoring can be affected by many factors. One of factors is smart students who write good sentences.

b. The comparison between objective test and subjective test:

   1) A subjective test or an objective test can be used to measure almost all of educational achievements.

2) An objective test or subjective test can be used to encourage students to learn, to understand the principles, compositions, or unification of ideas and application of knowledge in answering.

3) Subjective test requires students to string their idea up in their own words of their answer, whereas, objective test requires students to choose among some alternatives have been provided of a test item.

4) The subjective test consists of a more general question that expects a longer answer, whereas, objective tests usually consist of many more specific questions that require short answers.

5) The students spend most time to think and write a answer when dealing in subjective tests, and they spend most time to read, think, and choose a answer when facing the objective tests.

c. There are some kinds of objective tests; (1) true-false test, (2) matching test, (3) multiple choice test, (4) completion, (5) classification, (6) cause and effect.

d. There are some kinds of subjective test; (1) essay writing, (2) composition writing, (3) letter writing, (4) reading aloud.

The English second semester final test of the eighth grade students at SMPN 2 Semarang in the academic year 2017/2018 just provided multiple choice test and essay, however, the researcher just focused on multiple choice test which had been explained on limitation of the study in Chapter I. Therefore, the researcher would just like to focus on explanation of multiple choice test.

2. 2. 1. 6 Multiple choice Test

Multiple choice test is one of the most widely type used of summative test and reliable tool to evaluate students' learning performance. Multiple choice test can measure lower to higher cognitive processing of Bloom's taxonomy from remembering, understanding, applying, analyzing, evaluating, and creating (Kolte, 2015). Multiple choice test has certainly one correct answer (answer key) and the others as wrong answers (distractors). It can be question form or an incomplete statement. It is supported by Toksöz & Ertunç (2017, p. 142) "All of alternatives, there is a key which is defined as the most appropriate response to the stem, and the other alternatives are called distractors". For the alternatives, in multiple choice test of elementary school has three alternatives in grade 1, 2, 3, and four alternatives are for grade 4, 5, 6, junior high school has also four alternatives, and senior high school has five alternatives.

Teachers more like to make a test in multiple choice format as it is easy to prepare and practical to administer. Multiple choice test can also be checked easily and quickly. Therefore, it is easy for them to make a test in multiple choice format. Multiple choice test seem to be more reliable compared with other types of tests which is subjective tests.

Brown (2004, p. 56) and Shaban (n.d, p. 47-48), stated that multiple choice item divided into three parts:

a. The stem: the initial part of each multiple choice item is known as the stem. It can be a complete statement, an incomplete statement and question.

b. The correct choice, correct answer, correct option or key: the answer can be a word or a group of words.

c. The distractor (incorrect options or incorrect answer): the distractors can be two, three or four options.

2. 2. 2 The Characteristics of Multiple Choice Test

The characteristics of multiple choice test is item difficulty, item discrimination, alternatives (distractor and answer key), and reliability (Brown, 2004; Supranata; 2004; Suwarto, 2007, 2011, 2016).

2. 2. 2. 1 Item Analysis

Crocker & Algina (1986, p. 311) stated that item analysis is used to define the computation and examination of any statistical property of students to an individual test item. Thorndike (1949) as cited in Roid & Haladyna (1982, p. 215) stated that "the study of test items is desirable to produce effective test, since test depend on the characteristics of test items. The tests may be added or removed from test based on item characteristics". It means that the statements above is the traditional study of test item, while, in modern context, the main goal of item analysis dose not select items, but it reviews the quality of items in domain (Roid & Haladyna, 1982). The purpose of item analysis is to obtain information about objective test items used to indicate weakness of a item and to identify bad items (Suwarto, 2004). The items sometimes are easy, medium, or difficult to differ students of high group or low group. The writers' point of view of the item analysis above, it could be conclude that the item analysis is the way to identify quality of each item of objective test statistically based on the characteristics of multiple choice test.

To know the characteristics of multiple choice test can be done using qualitative (theoretical) and quantitative (empirical) analysis. Surapranata (2004, p. 1) stated that qualitative analysis is a review that is intended to analyze the questions in terms of technical, content, and editorial. Qualitatively, the test is said to be good test if it has the requirements from material, construction and language. According to Crocker & Algina (1986, p. 4), the quantitative can be done with two techniques: classic test theory (classical true-score theory) and the response theory of item (item response theory). This study, the researcher just focused on quantitative analysis with classic test theory.

Classic test theory is an easy theory that is quite useful in describing how the error in measurement which can affect the observation score. It is formulated systematically as well as in the long term. The important formulas of classic test theory are item difficulty, item discrimination, alternatives, reliability, and validity (Lababa, 2008). According to Suwarto (2004, p. 293), the item difficulty and the item discrimination are the most fundamental statistics in analyzing of an item.

a. Item Difficulty

The item difficulty is how difficult or how easy of item for students. To get item difficulty index, it can be computed from the percentage of pupils who answer correctly of an item compared by the total of examines taking the test item. "Any test item's difficulty can be calculated by noting the proportion of subjects in a representative sample that give correct responses” (Roid & Haladyna, 1982, p. 216). Suwarto (2004, p. 293) stated that it is also named proportional of correct (*p*). The item difficulty is the opportunity to answer a problem correctly at a certain level of ability which is usually expressed in the form of an index. Richard & Sheila (1999, p. 18) stated that the item difficulty has an index ranging from a low of 0.00 to a high of +1.00. The higher item difficulty indexes indicate easier items. For instance, a item has 0.00 item difficulty index, it means that there is no students answered correctly of the item. It indicates the item is difficult. Then, if the item difficulty has 1.00 item difficulty index, all of students will certainly answer correctly. It indicates the item is easy. In addition, Richard & Sheila (1999, p. 18) stated that “Item difficulty is a characteristic of the item and the sample that takes the test”. For example, a main idea question of a certain kind of texts will be easier for senior high school students, but difficult for elementary pupils. From the explanations above, it can be concluded that the item difficulty is the average score obtained by students on the item. Based on Richard & Sheila (1999, p. 18), it can be interpreted in the formula:

$$Difficulty = \frac{\sum who\ \text{answered an item corretly}}{\text{Total tested}}$$

The function of item difficulty is usually associated with the purpose of a test, for example, semester exam is used items that have medium category, for the purposes of selection used items that have high item difficulty, and for the purposes of diagnostic used items that have easy level (Rudyatmi & Rusllowati, 2017).

b. Item Discrimination

The item discrimination of each test item is the ability of an item to distinguish between the high and low student’s achievement and ranges between -1.00 to + 1.00 (Roid & Haladyna, 1982; Rudyatmi & Rusllowati, 2017). A item which has the higher the discrimination index, it indicates that the item is answered by upper group correctly. However, if lower group answer it correctly, it wil have a negative valued and is probably flawed. A negative discrimination index occurs if the item are too hard or poorly written which makes difficult to select the answer key for students. The negative sign shows that students who have low achievement can answer the test item, however, students who have high achievement cannot answer it. Thus, the item that has negative item discrimination index shows that students' ability is reversed Richard & Sheila (1999).

The item discrimination index can be showed in point biserial correlation or biserial correlation (Singh et al., 2014). The higher item discrimination index, it means that the item can distinguish between high and low student’s achievement in understanding material taught by their teacher. The function of the item discrimination is to detect individual differences among students. Rudyatmi & Rusllowati (2017, p. 96) stated that the advantages of analyzing item discrimination are improving each test item through the empirical data. Each test item can be identified whether the test item has to save, revised, or eliminated of test bank based on item discrimination index.

To determine item discrimination index are usually calculated by using correlation index, discrimination, and item alignment indices, but from three ways the most commonly used is the correlation index (Lababa, 2008). According to Crocker & Algina (1986, p. 317), there are four correlation indices used to calculate item discrimination: point biserial correlation, biserial correlation coefficient, phi coefficient, and tetrachoric correlation coefficient. The researcher used point biserial correlation to calculate item discrimination index statistically, because many teachers use it from view of (Rudyatmi & Rusllowati, 2017). There are two advantages using point biserial correlation coefficient that it makes more stable from sample to another sample and scoring is more accurate that each test item can differentiate some point differences in stability (Rudyatmi & Rusllowati, 2017). However, the researcher used biser to analyze the item discrimination in the ITEMAN version 3.00 program.

Therefore, by analyzing item discrimination of each item, we can improve the quality of each item through empirical data. Based on the item discrimination index, every item can be identify whether the item is saved, revised, or eliminated. We can know how far the item can differ students' ability in understanding material and students who does not understand material that their teacher taught before.

c. Alternatives

Semester test is certainly objective test and subjective test, the multiple choice questions as objective test and short answer as subjective test. This study, the researcher just focused on multiple choice test that has alternatives, because this study analyzed the English second semester final test of eighth grade which has four alternatives (A,B,C,D). One of the alternatives must be correct answer as answer key, and three of another alternatives must be false answer as distractors which have function and effectiveness to confuse students when they chose among the alternatives. By analyzing the alternatives, it can be known (1) The number of students who answer correctly, (2) a distractor that has many mistakes makes no one that choose the distractor. (3) the distractor interest for students who have low achievement (Surapranata, 2004).

1) Answer Key

Answer key is the only one option which has correct answer. It can be wrong or true, because it was made by a teacher. Therefore, the researcher also analyzed the answer key from English MGMP *Sub Rayon* 01. The answer key of each test item is analyzed by seeing key part of Prop. Endorsing that must be bigger than distractors' Prop. Endorsing, part of Biser must be bigger than distractors' Biser, part of point must be bigger than distractors' point, so the key could be good key. Otherwise, key needs to be checked and if key's prop endorsing, biser, point biser are smaller than distractors' prop endorsing, biser, point biser. It can point out that the key is incorrect or revised. (Suwarto, 2016).

2) Distractors

Distractors are the multiple choice response options which must be incorrect answer. Their function is students' common misconceptions or miscalculations, so there is distractors of alternatives in order to students are confused in choosing correct answer option. Distractors are analyzed to determine their relative usefulness of each test item. Distractor is said to be effective, if it is selected at minimum 5% (0.050) of the respondents. The respondents come from students who have low achievement. Distractors are said to be ineffective, if it is selected less than 5% of respondents. The ineffective distractors should be revised (Mutaqi, 2007; Suwarto, 2016). By using ITEMAN version 3.00 program, distractors can be known effective or not through Prop. Endorsing index.

2. 2. 2. 2 Reliability

Reliability of a test shows that the extent to which the measurement results can be trusted (Suryabarta, 1998; Suwarto, 2013). It means that a test must yield a dependable score. The use of the measuring instrument repeatedly will provide consistent results. It supported by Harrys and Valette (1992, p. 14) stated that "reliability means the stability of test scores. A test cannot measure anything well unless, and it measures consistently". According to Brown (2004, p. 20) stated that "a reliability test is consistent and dependable". Crocker & Algina, (1986, p. 105) argued that "consistency or reproduction of test scores is called reliability." Suhr (2003, p. 6) also stated "development of a measurement instrument is a complex process. Reliability assesses the accuracy and precision of the instrument". For example, if a teacher give the same test to the same students or matched students at the different time, the test should give similar result.

Based on the definition of reliability from another researchers and authors, it could be concluded that reliability is the level of accuracy, constancy, or stability. If a measure tool has high reliability or trusted, the test will be stable. It means that the test can be used to measure the students' performance. Reliability relates to the extent to which the tests are given from time to time. A test is said to be consistent from time to time to produce the same or relatively the same score.

The test can be unreliable may be some factors. According to Mousavi (2002) as cited in Brown (2004, p. 21), some factors that make the test to be unreliable are fluctuations of the students, in scoring, in test administration, and in the test itself. First, student-related reliability, test can be reliable or unreliable based on the condition of students, for instance, if the students are in a bad day, illness, anxiety, or other physical, it will impact for their observed test score that deviate their true score. The test can be unreliable. Second, rater reliability, human error, subjectivity, and bias may enter into the scoring process. Inter-rater reliability occurs when two or more scores are inconsistent scores of the same test because the teacher has lack of attention to scoring criteria, inexperience, inattention, or preconceived biases. The two scores are not applied the same standards. Third, test administration reliability comes from the condition of classroom, such as, the lighting in different parts of the rooms, variations in temperature, and the condition of students' desks and chair. Fourth, test reliability, the test itself can impact measurement errors. For example, if a test item is too long, students are able to need a long time to think and write, thus, they are able to be loss time and they may answer it incorrectly.

The reliability index ranges from 0 - 1. A test is said reliable if the reliability index is up to 0.700. The higher of reliability coefficient of a test that close to index 1, the higher accuracy of the test. A reliability coefficient of 1 indicates perfect reliability (Rudyatmi & Rusllowati, 2017; Roszkowski & Spreat, 2011).

There are four methods to estimate the reliability index (Nugiyantoro et al., 2002; Harrys & Valette, 1992):

a. Test-retest

Test-retest is the same two test which is done by same group in twice time. The time should not be too close, for instance, the test is held every two weeks or once a month. the result of them will be correlated with product-moment formula. The index (r) will suit with table of critical values r. If the r is ≥ 0.050, the r will significant. Therefore, the test is reliable.

b. Parallel test

Parallel test is a group doing two different tests but the both of different tests has same grid competency (parallel). It means that first test and second test must have the same level of components. The parallelism of the tests is latticework of making test. Their data result will be correlated with product-moment formula. The index (r) will suit with table of critical values r. If the r is ≥ 0.050, the r will significant. Therefore, the test is reliable.

The test-retest and parallel test are still hard to students and researcher because students must do two tests, and the researcher must make two tests. There is easier technique that is Internal consistency technique.

c. Internal consistency technique

To estimate reliability index with internal consistency technique is held to focus on test items. All items is N which is made a test. This method does not need two tests to obtain data. It indicates that this technique is more efficient than two techniques above. According to Nugiyantoro et al. (2002, p. 323) and Riduwan (2007, p. 102), there are four formula to calculate relebility index of a test. They are Spearman Brown, Kuder-Richardson 20 (KR-20), Kuder-Richardson 21 (KR-21), Anova Hoyt, and Alpha Chronbach. However, the researcher used Alpha Chronbach formula because output of ITEMAN version 3.00 used Alpha to point reliability index out.

2. 3 **Theoretical Framework**

Test is one of instruments to measure students' performance that consists of items that must be answered by them. According to Suwarto (2013, p. 9) the types of test are placement test, diagnostic test, formative test, and summative test. The item types are objective and subjective (Rudyatmi & Rusllowati, 2017, p.23-49). the kinds of objective test are true-false test, matching test, multiple choice test,completion, classification, and cause and effect. The kinds of subjective test are essay writing, composition writing, letter writing, and reading aloud

The researcher analyzed the summative test that was English second semester final test which consists of objective test and subjective test. However, the limitation of this study just focused on multiple choice test which is objective test. The answer will be objective for its correction because the multiple choice test certainly has one correct answer. Therefore, the researcher analyzed the characteristics of multiple choice test. According to Supranata (2004, p.10), the characteristics of multiple choice test are item difficulty, item discrimination, alternatives (answer key and distractors), and reliability. The framework can be presented in diagram form as follows:

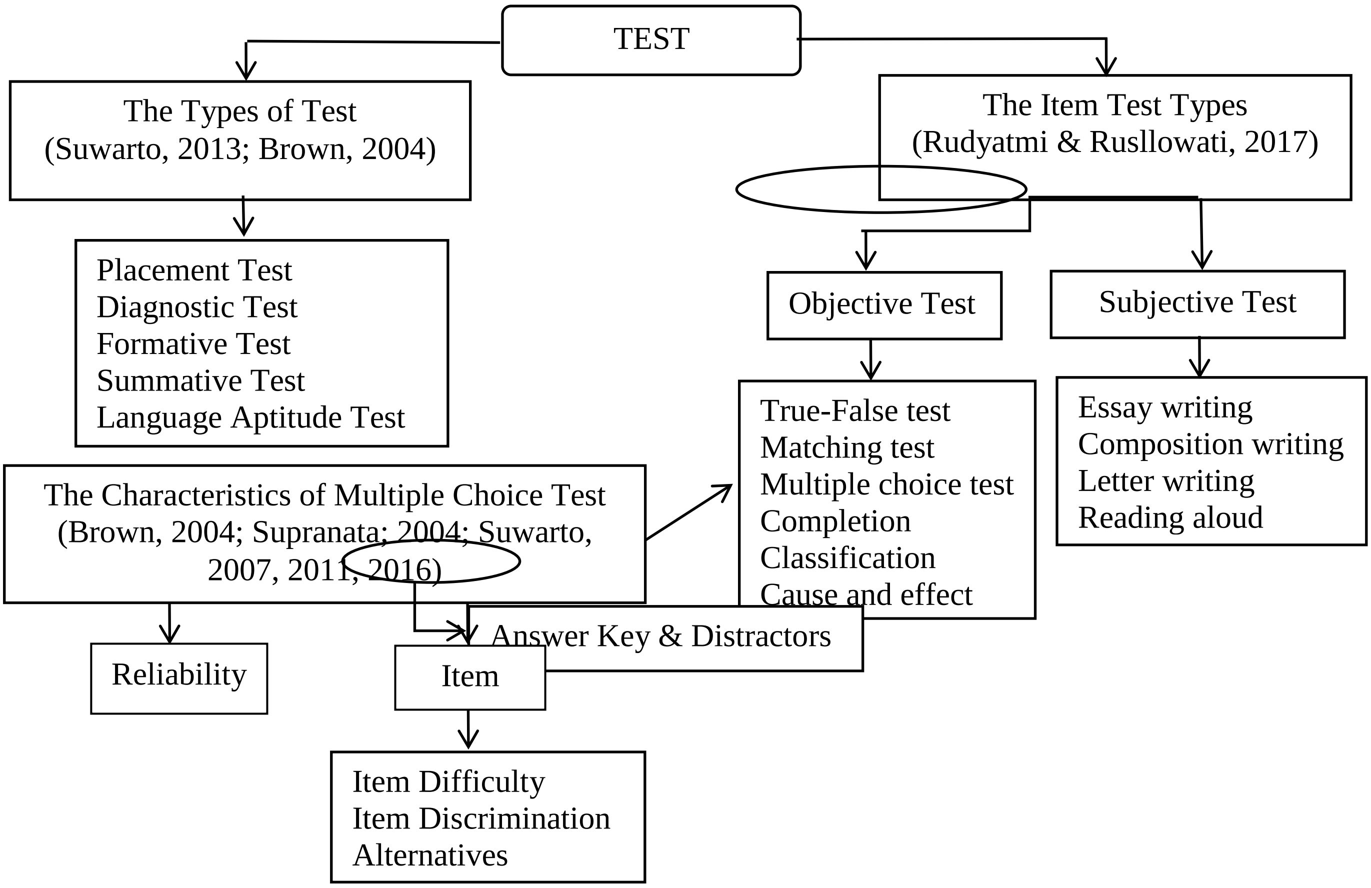


Figure 1 Theoretical Framework

CHAPTER V
CONCLUSIONS AND SUGGESTIONS

Chapter five provides some conclusions and suggestions for this study. The conclusions were based on the previous chapter, regarding the findings and discussions. It was also useful for readers because this part was the summary of this study. Furthermore, the suggestions were useful for a teacher as a test maker.

5. 1 Conclusions

Based on the data analysis of the characteristics of the English second semester final test for eighth grade students at SMPN 2 Semarang in the academic year 2017/2018, the researcher had concluded five main points from the fourth chapter.

1. The item difficulty index of the test had ranged from 0.117 until 1.000. The lowest item difficulty index was 0.117 (item 26) and the highest item difficulty index was 1.000 (item 11 and item 30). The percentage comparison of easy item: medium item: difficult item was 57.5% : 30% : 12.5%.
2. The item discrimination index of the test had range -9.000 until 1,000. The lowest item discrimination index was -9.000 (item 11, item 30) and the highest item discrimination index was 1.000 (item 13). The percentage comparison of bad item: sufficient item: good item: excellent item was 22.5% : 22.5% : 42.5% : 12.5%.
3. There were 65 ineffective distractors of 33 items which had to be revised and 55 effective distractors of 7 items which had to be saved in test bank. The percentage comparison of ineffective item: effective item was 54.17%:

45.83%. Then, the answer keys which must be cross checked were 2 item (item 3 and item 38).

4. The reliability of the test was 0,717. It meant that this test was reliable because the index was up to 0.700.

5. 2 Suggestions

Based on what the researcher described in the data analysis and interpretation, a test had to be analyzed its characteristics especially for the multiple choice from of this English second semester final test before the test was fulfilled by students. The test-maker should also pay attention to the characteristics of multiple choice test which are the item difficulty, item discrimination, alternatives (distractors and answer key), and reliability by trying out the test before the test was distributed to students. After they analyzed the characteristics, they could revise, edit, or delete the bad item. If the test items were very good, the reliability index of the test would automatically increase. It meant that the test had very high reliability. It would minimize the smallest possible error scores in the test to measure students' real achievement or their true score, so the test would be a trusted test and the students' achievement measuring would be more accurate. Not only it was the students' achievement, but also it helped easy for the teachers to diagnose their teaching learning, media, or method, thus, the teachers could improve them to succeed of their teaching leaning. They also could encourage a motivation for lower student to study hard, and give them a extended class or remedial to improve their students' knowledge.

REFERENCES

Allen, M. J. & Yen, W. M. (1979). *Introductioon to Measurement Theory.* Monterey: Brooks Cole Publishing Company.

Bernasela. (2014). An Analysis on English Summative Test Items. Tanjung Pura University.

Brown, H. D. (2004). *Language Assessment: Principle and Classroom Practice*. United States of America: Pearson Education.

Brown, J. D. (2001). Point-biserial Correlation Coefficients. *JALT Testing & Evaluation SIG Newsletter, 5* (3), 12–15.

Cambridge advanced learner’s dictionary (3nd ed.). (2008). Cambridge University Press.

Chauhan, P. R. (2013). Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot. Biomirror, 4 (6), 1-4 Retrieved from: www.bmjournal.in

Chellamani, K. & Boopathiraj, C. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in The Test for Research in Education. *International Journal of Social Science & Interdisciplinary Research 2* (2), 189-193. Retrieved from: indianresearchjournals.com

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College.

Djemari, M. (2004). *Penyusunan Tes Hasil Belaja.* Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Harrys & Valette. (2003). Principles Language Testing I. New York: Mc Graw Hill.

Haryudin, A. (2015). Validity and Reliability of English Summative Tests at Junior High School in West Bandung. *Jurnal Ilmiah UPT P2M STKIP Siliwangi, 2* (1), 77-90.

Hayat, B., Pranata, S. S. & Suprananto. (1997). *Manual* Item and TEst Analysis (ITEMAN). Jakarta: Pusat Penelitian dan Pengembangan Sistem Pengujian, Balitbang Dikbud.

Hidayati, A. D. (2009). The Analysis of Validity, Reliability, Discrimination Power and Level of Difficulty of First Mid-Term Test in The Case of The Eighth Grade Students of SMP 33 Semarang. A final project: Semarang State University.

Hughes, A. (2003). *Testing for Language Teacher*. Cambridge: Cambridge University Press.

Jandaghi, G. (2011). Assessment of Validity, Reliability and Difficulty Indices for Teacher-built Physics Exam Questions in First Year High School. *Arts and Social Sciences Journal*. ASSJ-16. E-ISSN: 21516200. Retrived from: http://astonjournals.com/assj

Kartowagiran, B. (2009). *Pengantar Teori Tes Klasik (TTK). Makalah Pascasarjana UNY dan DIinas Pendidikan Prov DIY.*

Khoshaim, H. B. & Rashid, S. Assessment of the Assessment Tool: Analysis of Items in a Non-MCQ Mathematics Exam. *International Journal of Instruction.International Journal of Instruction*, *9* (1), 119- 132. Retrieved from www.e-iji.net

Kolte, V. (2015). Item Analysis of Multiple Choice Questions in Physiology Examination. *Indian Journal of Basic and Applied Medical Research, 4(4),* 320-326 .

Lababa, D. (2008). *Analisis Butir Soal dengan Teori Tes Klasik*: *IQRA' Journal, Volume 5*, 29-37.

Lai, E. R. (2011). Motivation: A Literature Review. Research Report. Retrieved from http://www.pearsonassessments.com/research

Masruroh, H. Z. (2014). An Item Anaalysis on English Summative Test for Second Grade Students of MAN Tulungagung 1 in Academic Year 2013/2014. A Script: State Islamic Institute Tulungagung.

Mulianah, S. & Hidayat, W. (2013). *Pengembangan Tes Berbasis Komputer. Kuriositas*, *2*(6), 27- 43.

Mutaqi. (2007). Analisis Butir Soal Terhadap Instrumen Evaluasi Kegiatan Diklat. *Materi Workshop Direktur Diklat di UDIKLAT PT PLN (PERSERO) Semarang*, 9 April 2007.

Nugiyantoro, B., Gunawan, Marzuki. (2002). *Statistik Terapan Untuk Penelitian Ilmu-Ilmu Sosial*. Bulaksumur, Yogyakarta: Gadjah Mada UNiversity Press.

Pascual, G. R. (2016). Analysis of The English Achievement Test for ESL Learners in Northern Philippines. International Journal of Advanced Research in Management and Social Sciences, 5 (12),1-5. retrieved from www.garph.co.uk

Putri, N. S. (2015). An Analysis of English Semester Test Items based on The Criteria of A Good Test for The First Semester of The First Year of Smk Negeri 1 Gedong Tataan in 2012/2013 Academic Year. A Script: Lampung University.

Putri, Y. F. D. R. (2009). Analysis of Teacher-Made English Final Second Semester Test For The Year Eleven Students of SMAN 1 Ambarawa in The Academic Year of 2008/2009 based on the Representativeness of Content Standard. A Script: Universitas Negeri Semarang.

Raharja, N. S. (2014). *Analisis Butir Soal Ujian Akhir Sekolah Produktif Pemasaran Kelas Xii Pemasaran SMK Negeri 9 Semarang*. Economic Education Analysis Journal, 3(3), 564-569. Retrieved from http://journal.unnes.ac.id/sju/index.php/eeaj

Richard, J. & Sheila, C. (1999). *Item Analysis for Criterion-Referenced Tests*. New York: Research Foundation of SUNY/Center for Development of Human Services.

Riduwan. (2007). *Belajar Mudah Penelitian untuk Guru, Karyawan dan Peneliti Pemula*. Akdon. Bandung: Alfabeta.

Roid, G. H. & Haladyna, T. M. (1982). *A Technology for Test-Item Writing*. London: Academic Press, Inc.

Roszkowski, M. J., & Spreat, S. (2011). Issues to consider when evaluating “tests”. In *Financial planning and counseling* scales, 13-31. Springer New York.

Rudyatmi, Ely & Rusllowati, A. (2017). *Evaluasi Pembelajaran.* Semarang: Faculty of Mathematics and Science Unnes.

Rusmiana, F.D. (2015). The Test Item Analysis of 1st Semester Final Test of The Accounting Theory for Vocational Education: Case Study of SMK YPKK 1 Sleman Academic Year of 2014/2015. A Thesis: Yogyakarta State University.

Sa’adah, N. (2017). The Analysis of English Mid-Term Test Items based on the Criteria of a Good Test at the First Semester of the Eighth Grade Students Of Mts. Mathalibul Huda Mlonggo In The Academic Year Of 2016/2017. *Journal Edulingua, 4* (1),45-58.

Saleh, M. (2013). *Introduction to Linguistic and Educational Research*. Semarang: Faculty of Languages and Arts Unnes.

Saputra, R.W. (2015). The Comparison Between the Second Mid-Term English Tests for the Seventh Gradersmade by the State and Private School Certified English Teachers. *Journal of English Language Teaching, 4* (1), 1-5 ISSN 2252- 6706 Retrieved from: http://journal.unnes.ac.id/sju/index.php/elt

Setiyana, R. (2016). Analysis of Summative Tests for English. *English Education Journal, 7*(4), 433-447

Shaban, A. S. (n.d). A comparison between Objective and subjective tests. A paper.

Shomami, A. (2014). An Item Analysis of English Summative Test. A Script: Syarif Hidayatullah State Islamic University.

Singh, J. P., Kariwal P., Gupta S.B., & Shrotriya V.P. (2014). Improving Multiple Choice Questions (MCQs) through item analysis: An assessment of the assessment tool. *International Journal of Sciences & Applied Research, 1(2),* 53-57. Retrived from: www.ijsar.in

Sudijono, A. (2011). Pengantar Evaluasi Pendidikan. Jakarta: Rajawali Pers.

Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature, 3* (2), 22-38. P-ISSN: 2460-0938. E-ISSN: 2460-2604.

Suhr, D. (2003). reliability, exploratory & confirmatory factor analysis for the scale of athletic priorities. Retrieved from: http://www2.sas.com/proceedings/sugi28/274-28.pdf

Surapranata, S. (2004). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Test.* Bandung: PT Remaja Rosdakarya.

Suruchi & Rana S. S. (2014). Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology. *Indian Journal of Research, 3* (6), 56-58. ISSN - 2250-1991.

Suwarto. (2004). *Analsis Item dalam Pembutan Tes. Education Jounal,13*(3), 289-300.

Suwarto. (2007). *Tingkat Kesulitan, Daya Beda, dan Reliabilitas Tes Menurut Teori Tes Klasik. Journal Pendidikan, 16* (2), 166-178.

Suwarto. (2011). Teori Tes Klasik dan Teori Tes Modern. *Widyatama*, *1*(20). 69-78.

Suwarto. (2013). *Pengembangan Tes Diagnostik Dalam Pembelajaran*. Yogyakarta: Pustaka Pelajar.

Suwarto. (2013). *Tingkat Kesulitan, Daya Beda, dan Realibilitas Tes Ujian Seleksi Mahasiswa Baru Universitas Veeran Bangun Nusantara Sukoharjo*. National Seminar on Science Education: p. 652- 658.

Suwarto. (2016). Karakteristik Tes Biologi Kelas 7 Semester Gasal. *Jurnal Penelitian Humaniora*. *17*(1), 1-8.

Suwarto. (2018). *Statistik Pendidikan*. Yogyakarta: Pustaka Pelajar.

Zubairi, A. M. & Kassim, N. L. A. (2006). Classical And Rasch Analyses Of Dichotomously Scored Reading Comprehension Test Items. *Malaysian Journal of ELT Research 2*, 1-20. Retrieved from www.melta.org.my