A DESCRIPTIVE STUDY ON THE QUALITY OF ENGLISH FINAL TEST AT THE FIRST SEMESTER OF 12th GRADE STUDENTS OF SMAN 1 KEDUNGWARU IN ACADEMIC YEAR 2014/2015

THESIS

By:

ATIK LAELATUS SHOFIYAH

NIM: 3213113052

ENGLISH EDUCATION DEPARTMENT

FACULTY OF TARBIYAH AND TEACHER TRAINING

STATE ISLAMIC INSTITUTE OF TULUNGAGUNG

2015

DESCRIPTIVE STUDY ON THE QUALITY OF ENGLISH FINAL TEST AT THE FIRST SEMESTER OF 12th GRADE STUDENTS OF SMAN 1 KEDUNGWARU IN ACADEMIC YEAR 2014/2015

THESIS

Presented to Faculty of Tarbiyah and Teacher Training
State Islamic Institute of Tulungagung
In partial of fulfillment of the requirements for the degree of Sarjana Pendidikan Islam (S. Pd.I) in English Education Department

By:

ATIK LAELATUS SHOFIYAH

NIM: 3213113052

ENGLISH EDUCATION DEPARTMENT

FACULTY OF TARBIYAH AND TEACHER TRAINING

STATE ISLAMIC INSTITUTE OF TULUNGAGUNG

2015

ADVISOR'S APPROVAL SHEET

This is to certify that a thesis entitled " A Descriptive Study on the Quality of English Final Test of the 12th Grade Students of SMAN 1 Kedungwaru at the First Semester in Academic Year 2014/2015" written by Atik Laelatus Shofiyah, Student Registered Number of 3213113052 has been approved by the thesis advisor for further approval by the Broad of Examiners.

Tulungagung, May 19th, 2015

Advisor

Arina Shofiya, M.Pd
NIP. 19770523 200312 2 2002

Approved by

The Head of English Education Department

Arina Shofiya, M.Pd
NIP. 19770523 200312 2 2002

BOARD OF THESIS EXAMINERS' APPROVAL SHEET

This is to certify that a thesis entitled " A Descriptive Study on the Quality of English Final Test of the 12th Grade Students of SMAN 1 Kedungwaru at the First Semester in Academic Year 2014/2015" written by Atik Laelatus Shofiyah , Student Registered Number of 3213113052 has been approved by the Broad of Examiners as the requirement for the degree of Sarjana Pendidikan Islam in English Education.

Tulungagung, 19th May 2015

Board of Thesis Examiners

Chair,	Secretary,
<u>Nanik Sri Rahayu, M. Pd</u> **NIP.19750707 200312 2 002**	**<u>Arina Shofiya, M.Pd</u>** **NIP. 19770523 200312 2 002**

Main Examiner

Ida Isnawati, M. Pd
NIP. 19780816 200604 2 002

Approved by

The Dean of Faculty of Tarbiyah and Teacher Training

Dr. H. Abd. Aziz, M.Pd.I
NIP. 19720601 200003 1 002

MOTTO

Don’t ever ask if you never try

Your success depends on your struggle

Nobody will help you, yet you yourself and only you

DEDICATION

After finishing this thesis, I want to dedicate this thesis to:

1. My parents, H. Abbas Syamsuddin and Siti Khasanah who always pray for my success and who always give motivation to me
2. My beloved advisor, Arina Shofiya who always guided and gave me unforgettable knowledge
3. My close friends, Ely, Sugi, Aris, Dianti, Binti, Tembem, Lutfi, Ardiana, Fatma, Hida, Firdha, Diana, Amik, Zulaiha, Amalia who accompanied and supported me during doing this research
4. All members of TBI-8B whom I love

DECLARATION OF AUTHORSHIP

The undersigned below

Name : Atik Laelatus Shofiyah

Place, date of birth : Blitar, August 4th, 1992

Address : Ds. Mangunan RT/RW : 01/02 Udanawu Blitar

Department : Islamic Education Department (Tarbiyah)

Program : English Department

States that this thesis is truly my original work. It does not incorporate any material previously written or published by another person except those as indicated in quotation and bibliography. Due to the fact, I am the only person responsible for the thesis. If a later time it is found that this thesis is a product of plagiarism, I am willing to accept any legal consequences that may be imposed to me.

Tulungagung, April 29, 2015

Atik Laelatus Shofiyah
NIM. 3213113052

ABSTRACT

Shofiyah, Atik Laelatus. Student Registered Number 3213113052. 2015. *"A Descriptive Study on the Quality of English Final Test at the first semester of 12th Grade Students of SMAN 1 Kedungwaru in Academic Year 2014/2015".*Thesis.English Education Program. State Islamic Institute (IAIN) of Tulungagung. Advisor: Arina Shofiya, M.Pd

Keywords: final test, validity, reliability, difficulty level, discrimination power, distractor efficiency

One of the essential parts of teaching and learning process is evaluation because conducting an evaluation can give any information about the students, and also the effectiveness of teaching and learning process itself; and the information taken, later, can be used to the improvement of teaching and learning program. One of the instruments in doing evaluation in teaching and learning program is a test. The result of the test will represent the students' language proficiency of learning language, thus it is necessary to create a good test. A test is considered to be good if it fulfill the characteristics of a good test; validity, reliability, difficulty level, discrimination power, and distractor efficiency if the test is in the form of multiple-choice test.

The formulation of the research problem was how is the quality of English final test of the 12th grade students at the first semester made by SMAN 1 Kedungwaru in term of its validity, reliability, level of difficulty, discrimination power, and distractor efficiency?

The purpose of this study was to present the quality of the English final test of the 12th grade studentsmade by SMAN 1 Kedungwaru in term of its validity, reliability, difficulty level, discrimination power, distractors efficiency.

The research method applied in this research were: 1) the research design in this study was descriptive with quantitative approach, 2) the population of this study was the English final test; test-package A and B; and students' answer sheets of the 12th grade students at the first semester, 3) the sample was the tenglish final 40 students' answer sheets of the 12th grade students which was taken randomly, 4) the research instrument was documentation, and 5) the data analysis method was test item analysis.

The findings of this study showed that both test-packages were lack of content and construct validity. In term of the content validity, both tests- packages did not fully test all materials stated in the syllabus, furthermore, one of the skills of language was not tested at all, listening. Related to the construct validity, some of the techniques of testing used to test language skills were not relevant to the language testing theory, especially writing and speaking because the test was in the form of multiple-choice while these skills need practicing in order to evaluate them. Then, one of the test-packages categorized to have low reliability with the coefficient reliability of 0.48, and another one was high with the coefficient reliability of 0, 72. The analysis on the level difficulty of both test-packs showed that the percentage of the easy items of test-package A was 72.5%, and 60% for test-package B; fair items was 17.5% of test-package A and 27.5% of test-package B; and difficult items was 10% of test-package A and 7.5% of test-package B. It means that both test-packages were too easy for the students. Next, for the discrimination power of both test-packages were 20% of excellent test items for test-package A and 12.5% for test-package B; 5% of good test items for test-package A and 17.5% for test-package B; 70% of poor test items for test-package A and 62.5% for test-package B; and 5% of very poor items for test-package A and 7.5% for test-package B. It means that both-test-packages could not really discriminate the students. In line with the discrimination power, the effective of the distractor analysis for both test-packs also showed bad result in which the distractors were dominated by the omit distractors with the percentage of 83.125% for test-package A and 65.385% for test-package B. Omit distractor means that the distractors must be removed or revised totally.

ABSTRAK

Skripsi dengan judul *“A Descriptive Study on the Quality of English Final Test at the first semester of the 12th Grade Students of SMAN 1 Kedungwaru in Academic Year 2014/2015”* disusun oleh Atik Laelatus Shofiyah. 3213113052. Jurusan Pendidikan Bahasa Inggris di IAIN TULUNGAGUNG tahun akademik 2015, dan dibimbing oleh Arina Shofiya, M.Pd.

Kata Kunci: ujian semester, validitas, reliabilitas, tingkat kesukaran, daya pembeda, keefektifan pengecoh.

Evaluasi merupakan salah satu hal yang terpenting dalam proses belajar mengajar, karena dari evaluasi tersebut guru dapat memeperoleh banyak informasi tentang siswa, and juga keefektifan proses belajar mengajar yang berlangsung di kelas yang nantinya informasi tersebut akan dapat digunakan untuk perkembangan program pengajaran. Salah satu cara untuk melakukan evaluasi adalah dengan menggunakan test. Hasil dari tes tersebut nantinya akan digunakan sebagai tolok ukur untuk mengetahui sejauh mana pencapaian siswa selama proses belajar. Oleh sebab itu, sangat perlu bagi para pembuat test untuk menciptakan test yang baik. Sebuah test dikatakan baik apabila telah memenuhi karakteristik dari tes yang baik, yaitu: validitas, reliabilitas, tingkat kesukaran,m daya pembeda, dan keefektifan pengecoh apabila test tersebut berbentuk test pilihan ganda.

Rumusan masalah yang diangkat dalam penelitian ini adalah bagaimanakah kualitas soal ujian semester satu kelas 12 SMAN 1 Kedungwaru Tulungagung dalam hal validitasnya, reliabilitasnya, tingkat kesukarannya, daya pembedanya dan keefektifan pengecoh?

Tujuan penelitian ini adalah untuk mendeskripsikan informasi tentang kualitas butir soal ujian semester 1 kelas 12 SMAN 1 Kedungwaru Tulungaung dalam bidang validitasnya, reliabilitasnya, tingkat kesukarannya, daya pembeda dan keefektifan pengecoh.

Metode yang digunakan dalam penelitian ini adalah: 1) penelitian ini berbentuk deskriptif dengan menggunakan pendekatan kuantitatif, 2) populasi dalam penelitian ini adalah soal ujian semester ganjil yang terdiri dari dua jenis soal A dan B, dan lembar jawaban siswa, 3) sampel dalam penelitian ini adalah soal ujian semester tersebut dan 40 lembar jawaban siswa kelas 12 yang dipilih secara acak, 4) instrumen yang digunakan dalam penelitian ini adalah dokumentasi, 5) metode analisa data menggunakan analisis butir soal

Hasil dari penelitian ini menunjukkan bahwa kedua tes A dan B masih memiliki kekukarangan dalam bidang validitas isi dan konstruknya. Dalam hal validitas isi, kedua tes A dan B tidak sepenuhnya mengujikan materi yang tercatat di silabus, bahkan salah satu skill, yaitu skill mendengar, yang seharusnya diujikan tidak diujikan sama sekali. Sementara itu, dalam bidang validitas konstruk, kekurangannya terletak pada teknik yang digunakan untuk menguji skill menulis dan berbicara dimana kedua skill ini tidak seharusnya diujikan melalui soal pilihan ganda, melainkan ujian praktik. Reliabilitas dari kedua tes juga berbeda, salah satu tes memiliki tingkat reliabilitas yang tinggi dengan koefisien reliabilitas 0.72 sementara yang lain rendah yaitu 0.48. Tingkat kesulitas dari kedua tes adalah 72.5% soal mudah bagi tes A dan 60% bagi tes B; 17.5% soal cukup dari tes A dan 27.5% dari tes B; 10% soal sulit dari tes A dan 7.5% dari tes B. Sedangkan untuk daya pembedanya adalah 20% soal dari tes A dan 12.5% dari tes B merupakan soal dengan daya pembeda sangat bagus, 5% dari tes A dan 17.5% dari tes B memiliki ndaya pembeda yang cukup baik, 70% tes A dan 62.5% tes B buruk, dan 5% tes A dan 7.5% tes B sangat buruk Keefektifan pengecoh berbanding lurus dengan daya pembeda dimana 83.125% pengecoh dari tes A dan 65.385% pengecoh dari tes B merupakan pengecoh yang sangat buruk karena tidak dipilih sama sekali oleh siswa, sehingga pengecoh ini harus dihapus.

ACKNOWLEDGEMENT

In the name of Allah SWT, The Most Beneficient and The Most Merciful. All praises are to Allah SWT for all His blessings so that the writer can accomplish this thesis. In addition, may Peace and Salutation be given to the prophet Muhammad SAW who has taken all human being from the Darkness to the Lightness.

The writer would like to express the genuine gratitude to:

1. Dr. H. Abd. Aziz, M.Pd.I., the Dean of Faculty of Tarbiyah and Teacher Training of IAIN Tulungagung for his permission to write this thesis.
2. Arina Shofiya, M.Pd., the Head of English Education Department and also the writer's thesis advisor who has given me some insight, her invaluable guidance, suggestion, and feedback so the writer can accomplish this thesis.
3. Drs. Harim Soejatmiko, MM, the headmaster of SMAN 1 Kedungwaru in academic year 2014/2015 for the cooperation as the sample of this research.
4. My collaborative teacher, Dra. Lilik Anjarwati who has given me the valuable help and support during the study.
5. Writer's countless gratitude is given to all persons given their helps and support to accomplish this thesis.

The writer realizes that this research is far from being perfect. Therefore, any constructive criticism and suggestion will be gladly accepted.

Tulungagung, April 29, 2015

The writer

Atik Laelatus Shofiyah

TABLE OF CONTENTS

CHAPTER III RESEARCH METHOD

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

This introduction chapter presents background of the study, statement of the research problem, objective 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 essential part of teaching and learning process. The importance of evaluation cannot be replacable. It is important as an instrument of the school system, to the teacher, learner, parent and administrator for the improvement of instruction. It gives a huge information about the students for the contribution of teaching and learning program. Bumagat (2004: 5) states that teaching, learning and evaluation are three interdependent aspects of the educational process. Therefore, evaluation is an indispensable part of teaching-learning process.It is a means of determining the effectiveness of teaching methodologies, instructional materials and other elements affecting the teaching-learning situation. The aim of evaluation itself is to evaluate students' achievement and students' progress in teaching and learning process.

Through evaluation, pupils' achievement, interest, success, difficulty and instruction can be assessed properly. The result of evaluation can be used as a benchmark for instructional enhancement. The purpose of evaluation in teaching and learning especially in the language teaching and learning program is that to know the students' language mastery level which consists of four language skills; speaking, listening, reading, and writing; and the language components; pronunciation, vocabulary, and grammar (Bumagat, 2004: 7). Thus, when the teachers plan to make or create an evaluation, the evaluation must cover all of those skills and components because the result of the evaluation will be used as the representative of the students' achievement in mastering the material.

There are many kinds of evaluation, but the most commonly used is test. In order to evaluate students' ability in understanding the material have been taught by the teachers, the teachers usually give some questions related to the material have been taught in the form of a test. Usually, teachers conduct a test in the mid-term and in the final semester after all of the material taught. However, some teachers may also give a test on the last chapter in order to know how far the students understand the material for one chapter. And it is also possible for the teacher to give daily test or quiz to the students in order to know the students'progress in every meeting.

In doing evaluation, what kind of test and when the test will be conducted depends on what teacher wants; whether the teacher wants to know students' progress or they want to know students' final achievement. As stated by Hughes (1989: 10) in (Allison, 1999: 80-81) that it is helpful to distinguish further between "final" achievement tests and "progress" achievement tests. Final test can be based on either (a) a detailed syllabus plus content, such as the actual books and other materials set; (b) the objectives of the course,while progress tests tend to involve the actual materials in use during the course.

In creating a test, there are some components that test designer should consider. Hughes (1989:11) goes on to argue that to base test content on course objectives is much preferred; it will provide more accurate information about individual and group achievement to certain objectives determined by the teachers and it is likely to provide a more beneficial backwash effect on teaching.This statement sounds very reasonable, but however, the test designer must also consider about the students' language proficiency level.

In creating a test, the items of the test can be various; it can be in the form of multiple choice test, essay, or even oral test, but the most common form of items used by the teachers for junior or senior high schools are in the form of multiple choice test and essay.

The first form is multiple choice tests. This kind of test item is considered to be simpler than essay in the form of scoring method because teachers only count the correct answer. However, good multiple-choice test is not easy to make because the test designer must consider about its validity, reliability, discrimination power, index of difficulty and the last is the distractor efficiency. Thus, it is very necessary to create good test items for the students because the result of the test will be used as the representative of the students' ability so that the test items must be valid and reliable and it must also complywith the characteristics of a good test in term of the index of difficulty, discrimination power and distractor efficiency.

In addition, Allison (1999: 16) also stated that “Traditional testing concerns over reliability and validity—and, less technically, over test fairness – also remain relevant when informal tests and assessment procedures are to be used as a basis for major decision about individuals. Both teachers and learners may need to be satisfied, in the name of fairness, that a procedure assesses what it is claimed to assess (validity), and that it does so accurately (reliability).” In his statement, Allison emphasized the importance of validity and reliability in the test items; and this is what test designer must consider in creating a test paper.

The second form is essay test. By having an essay test, teachers will know how deep the students understand on the material because in the form of essay test, students will not be provided by the optional answers but they must create their own answer by their own language. So that, from the students’ answer, teachers will be able to know whether the students are really understand.

A problem arises when most teachers underestimate an evaluation of the test item in the english final test they have made, whereas, this evaluation is in fact so important for the teachers in order to know the quality of the test they made; whether it is already valid and fulfill the characteristics of good test or not. Analyzing test items include analyzing the validity, reliability, level of difficulty, discrimination power and distractor efficiency. By analyzing those aspects, teachers will know whether the test they made is already valid and reliable or not, whether the test is too easy or too difficult for the students, whether the test can discriminate the upper and lower students or not and teachers should make sure that the distractors they made are really able to distract students' answer because the better the test items constructed by the teacher is the more reliable the score of the students and the reliable score can be used as the representative of the students' ability.

In the previous study, Salwa (2012) conducted a research entitled "The Validity, Reliability, Level of Difficulty and Appropriatness to The Curriculum of English Test". The topic of her research is test item analysis. In her study, she compared the quality of the English final test of the first semester students grade V made by the English KKG of a ministry of education and culture and ministry of religion Semarang. The research design used was descriptive comparative with mix method. The finding showed that the qualities of both test-packs are balance, but then, in their qualitative aspects, the test-packs made by the Ministry of Education and Culture has better quality than another because the findings showed that there were some errors exist in the test-pack made by the Ministry of Religion.In this study, the researcher will not conduct a descriptive comparative research, yet a descriptive study with quantitative method.

In this study, the researcher took upper secondary level of education because this level is the highest level of education before enrolling to the university. When the graduated students of this level want to enroll to the university, the committee of the university will also consider about the score of the students during semesters; especially for the special enrolment such as PMDK. Therefore, it is very necessary for the teacher to create English final test which is valid and reliable in order to result on the reliable score. Reliable score can only be produced from a test which is based on the characteristic of good test.

The researcher had determined the school where the sample of this study taken; SMAN 1 Kedungwaru. The researcher chose SMAN 1 Kedungwaru because based on the data on Ministry of Education and Culture in Tulungagung in academic year 2014/2015 that SMAN 1 Kedungwaru is the best school in Tulungagung.Thus, the researcher was interested to analyze the test items used in the best school in Tulungagung. The english final test of the first semester in SMAN 1 Kedungwaru was held on Saturday December 13, 2014. The item of this test which has been carried out was never analyzed before. It means that the quality of the test items was never known whether it is valid and reliable or not. Thus, it is necessary to conduct an analysis on this item in order to know and to improve the quality of the test item itself.

Based on the explanation above, the writer was interested in conducting a research entitled“A DESCRIPTIVE STUDY ON THE QUALITY OF ENGLISH FINAL TEST AT THE FIRST SEMESTER OF THE 12th GRADE STUDENTS OF SMAN 1 KEDUNGWARU IN ACADEMIC YEAR 2014/2015”

B. Research Problem

According to the background of the study, the researcher formulated a research question; How is the quality of English final test at the first semesterof the 12th grade studentsof SMAN 1 Kedungwaru in term of its validity, reliability, level of difficulty, discrimination power, and distractor efficiency?

C. Objectives of the Research

According to the research problem that was defined previously, the purpose of this research is to present the quality of the English final test at the first semester of the 12th grade students of SMAN 1 Kedungwaru in term of its validity, reliability, difficulty level, discrimination power, distractors efficiency.

D. Significance of The Research

Related to the objectives of the study, this analysis was intended to seesome advantages as elaborated in some paragraphs below. There are threemajor significances that this study wants to contribute.

The first one is theoretical significance. This study may give basicunderstanding to the researcher and teachersthat assessmentand evaluation cannot be made and assumed only by basing on students' outer performance or guessing in some cases. They should know that the testi tems should be made to evaluate students'understanding and ability. Thus, the tests will be also useful to develop their professionalism as being aneducator.

The second one is practical significances. This study is beneficial for the test makers as additional reference in constructing and analyzing test items and also for other researchers as additional reference in conducting such kind of research in the future occasion.

The last one is pedagogical significance. This study provides English teachers especially Senior High School teachers with some meaningful and useful information of effective evaluation inteaching learning processand improvement in test making.

E. Scope and Limitation of the Research

The scope of this research covers validity, reliability, level of difficulty, discrimination power, and distractor efficiency of the english final test at the first semester students grade twelf in SMAN 1 Kedungwaru academic year 2014/2015. In this study, the researcher had analysis limitation in analyzing the validity of the test items. For analyzing the test items validity, the researcher limits the analysis only for content validity and construct validity and in this study, the researcher also cannot guarantee whether the students cheated or not during answering every item on the test.

F. Definition of Key Terms

There are several key terms that are used in this study. They are Validity, Reliability, Item Facility, Item Discrimination and Distractor.

1. Validity of a test is the most important principle of language testing. By far the most complex criterion of a good test is validity, the degree to which the test measures what it is intended to measure

(Brown, 2000:387). A valid test means that the test measured or tested what it should be measured or tested because a valid test will result the valid score.

2. Reliability of a test is determined mostly by the quality of the items, but it is also determined by the length of the test (Fulcher; 2010:57). Reliability refers to the consistency of score resulted from conducting one set of test twice to the same group and the result of the test should be similar or almost the same. If the test is similar, it means that the test is reliable and the score resulted from a reliable test is truly trusted.
3. Level of difficulty (Item Facility) is the extent to which an item is easy or difficult for the proposed group of test-takers. Arikunto (2012: 222) stated that Item facility (IF) is a statistical index showing the percentage of students who correctly answer a given item in the objective test. The higher this proportion, the lower the difficulty is.
4. Discrimination may be conceptually understood as the relationship between the difficulty of the test items and the ability of the examinees in answering the question. It is an index for determining differences among individual examinees; the upper and lower group; on the subject matter being assessed (Osterlind; 2002: 275).
5. Distractor is the optional answers made in the multiple choice test purposes to outwit the students' choice of the correct answer. Arikunto, (2012:233) defined distractor as the distribution of test-takers in choosing the optional answer (distractor) in multiple-choice questions.

CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter presents some references related to this study. They are Previous Studies, Language Testing and Assessment, Testing Langugae Skills and Components, Types of Assessment and Testing, and Test Item Analysis.

A. Previous Studies

This research refers to the previous study by Hanik and Fahru (2012) entitled "An Analysis of English Summative Test for 6th Grade Students in Three Public Elementary Schools in Udanawu Distric, Blitar Regency and Athiyah Salwa (2012) entitled "The Validity, Reliability, Level of Difficulty and Appropriatness to The Curriculum of English Test".

In the previous study conducted by Hanik Huzaimatul Husna and Fahrurrazy entitled " An Analysis of English Summative Test for 6th Grade Students in Three Public Elementary Schools in Udanawu Distric, Blitar Regency, Husna and Fahrurrazy intended to find out the quality of the English summative test for 6th garde students in three public schools in udanawu, Blitar in term of the test construction, content validity, reliability, level of difficulty, level of discrimination, and the effectiveness of distractors in descriptive evaluative research design. In this study, the researcher used mixed method in analyzing the data. The qualitative analysis method was used for evaluating the writing of the test construction, while the quantitative analysis method is used to evaluate the test items.

The finding of this study shows that first, the teachers generally know the principles to construct the three test format. The construction of wh-question, multiple-choice, and completion are excellent because the test construction fulfill principles of test construction, but based on the analysis the researcher finds some mistakes that should be revised by teachers. Second, based on the content validity, the materials being tested in the items do not cover all the basic competences in the School-Based Curriculum. Third, the reliability of wh-question and multiple-choice indicates that the overall test have high reliability, but for completion format has moderate reliability. Fourth, generally, the level of difficulty of each item format is fair. Fifth, the level of discrimination for all item formats is very good. The last, the distracters in multiple-choice format are mostly effective.

The content is also relevant to the previous study by Athiyah Salwa (2012) entitled “The Validity, Reliability, Level of Difficulty and Appropriatness to The Curriculum of English Test”. In her research, Athiyah investigated and compared the quality of the English final test of the first semester students grade V made by the English KKG of a ministry of education and culture and ministry of religion Semarang. In her study, the population was the English final test used in Elementary Schools in Semarang and the samples were English final tests of the first semester students grade V made by KKG of Ministry of Education and Culture; SDIT Al Kamilah; and Ministry of Religion Semarang; MI Darussalam. The research design used by Athiyah was descriptive comparative with quantitative and qualitative approach. She used descriptive because she wanted to present and describe the quality of both tests and she compared the test-packs because she wanted to know whether there was difference between those two test-packages or not. While the quantitative research design is used to identify the test items itself in term of its validity, reliability, level of difficulty and discrimination power. The finding showed that the qualities of both test-packages are balance, but then, in their qualitative aspects, the test-packs made by the Ministry of Education and Culture has better quality than another because the findings showed that there were some errors exist in the test-pack made by the Ministry of Religion

This research was little bit different in term of the data analysis. If the two previous researchers used mixed method in analyzing the data, in this research the writer only used quantitative research design in order to get the maximum result and to limit the research. In this research, the researcher usedEnglish Final Tests of the twelfth grade students of Senior High Schools in Tulungagung at the first semester made by SMAN 1 Kedungwaru. This study involved an analysis of the test items such as validity, reliability, item facility, item discrimination, and distractor efficiency analysis.

B. Language Testing and Assessment

A test is a method of measuring a person's ability, knowledge or performance in a given domain (Brown, 2000:384). In his definition, Brown wants to highlight that people's intelligence and achievement can be explored through testing. Some people may assume that the term testing and assessment is the same, however, those terms are actually so far different in term of the application, but the same in term of the purpose.

Alderson (1997:215-216) and other shave argued that "Testers have long been concerned with matters of fairness and that striving for fairness is an aspect ofethical behavior, others have separated the issue of ethics from validity, as anessential part of the professionalizing of language testing as a discipline". In short, it can be said that test is a part of assessment so that assessment is wider than test itself, while the term assessment can be understood as a part of teaching and learning process and both of them have the equal purpose; that is to know and evaluate students' strength and weaknesses. Thus, in the teaching and learning process, teachers should use both testing and assessment as a method in evaluating the students.

C. Testing Language Skills and Components

1. Testing Listening

An effective way of developing listening skill is through the provision of carefully selected methods. Such method is in many ways to that used for testing listening comprehension. Hughes (1989:134-135) states that testing listening involves testing macro and micro skills in listening. The macro skills of listening include; listening for specific information, obtained gist of what is being said, and following instruction. The micro skills of listening include level interpretation of intonation patterns and recognition of function of structures. Weir (1990: 57) suggested the techniques that are possibly used in testing listening:

a. Multiple Choice

   This technique may be considered as the simple technique of testing, however, this technique has disadvantages for testing and it is greater for listening testing; the test takers should listen to passage while reading the alternative options; thus it can disturb test takers focus.

b. Information Transfer Technique

   This technique is useful for 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 etc.

c. Dictation

   This involves the students to listen to dictated material which incorporates oral message typical of those and it might encounter in the target situations.

d. Listening Recall

In this technique, the students are given a printed copy of passage from which certain content words have omitted and the students should fill those omitted part.

e. Note Taking

This technique invites the students to take a note while listening to lecturer. This activity can be suite realistically replicated in the testing listening for some situations.

f. Recording and Live Presentation

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

2. Testing Speaking

The objective of teaching spoken language is the development of the ability to interact successfully in that language, and this involves comprehension as well as production (Hughes, (1989:101)). Consequently, test should elicit behavior which truly represent the students' ability and which can be scored validly and reliably. Here are the lists of the more useful and potentially valid techniques for testing speaking ability suggested by Weir (1990: 78-80):

1. Verbal Essay

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

2. Oral Presentation

   Students are asked to give a short talk on a topic which they have either been asked to prepare before and or have been informed shortly before test.

3. Free Interview

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

4. Controlled Interview

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

5. Information Transfer; Describing picture in 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. Before describing the picture, student is giving a few minutes for preparation.

6. Information Transfer; Question on a single picture

   The examiner asks the students a number of questions about content of picture, which they had studied.

7. Interaction Tasks

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

8. 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.

9. Imitation

   The students hear a series of sentences, then they should repeat each part of the sentences in turn.

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 process, 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 process involved in reading and the production of appropriate exercise and test materials to assist in the mastery of these processes.

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

- Scanning text to locate specific information
- Skimming text to obtain the gist
- Identifying stages to an argument
- Identifying examples presented in supporting sentences

While the micro skills underlying reading skills are:

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

Then, here 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 techniques that might be used to test reading as follows:

a. Multiple Choice Questions (MCQs)

   It is usually set out in such way that student is required to select the correct answer from the given options.

b. Short Answer Question

   It requires the students to write down specific answer in space provided on the question paper.

c. Cloze Test

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

d. C – Test

In 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. Then, the students complete the words on the test paper and an exact word scoring procedure is adapted.

e. Selective Deletion Gap Filling

In this technique, the constructor should use a “rationale cloze” selecting items for deletion based upon what is known about language.

f. Cloze Elide Test

It is a technique which is generating interest where words which do not belong are inserted into a reading passage and students have to indicate where these insertions are 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 pictures, and etc.

Hughes (1989:131) advised to obtain reliable scoring, error grammaticality, spelling or pronunciation should not be penalized, and if it is clear, the students have successfully performed the reading task which the items set.

4. Testing Writing

The best way to test students' writing is to get them to write directly. Therefore, indirect writing testing cannot possibly be 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 many kinds of writing abilities that we engaged in. another reason for the variety of writing tests in use is the great numbers of factors that can be evaluated in writing skill; 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, and occasion), and etc.

Weir (1990: 66) suggested the techniques to test writing as follows:

a. Editing Task

In this kind of test, students are given a text containing a number of errors of grammars, spelling and punctuation of the type noted as common by remedial teachers of the students in the target group and 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 to perform certain functional tasks required in the performance of duties in the target situation, here are some kinds of direct writing tests:

a. 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 sentences.

b. Controlled Writing Task

   It tests important skills which no other form of assessment can be sampled adequately. Omitting a writing task in a situation where writing task is an important feature of the students' real life needs might be severely lower of the validity of testing programs.

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

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

5. Testing Vocabulary

The purpose of vocabulary testing is to measure the comprehension and production of words used in speaking or writing. The specifications of vocabulary achievement test should be based on all items presented to the students in vocabulary class. There are four general kinds of vocabulary tests that are presented by Madsen (1983: 12-30) as follows:

a. Limited Response

This kind of technique is very suitable for children and beginning level adults because they don’t have to know how to read or write, in fact, they don’t even have to know how to speak. Here is the illustration of this technique:

- Write out five commands that a student can perform individually by moving about the room, and five command that he can perform while sitting.
- Write out five commands or questions that a student can respond individually by pointing to a picture that you have prepared.
- Use the picture from activity 2 and prepare five requests that require students to follow instruction by drawing.
- Use original line drawing or picture from your students’ text showing activities, then prepare five vocabulary questions that require short answer. Then supply sample answer to be chosen by your students.

b. Multiple – Choice Completion

It is a good vocabulary test type for students who can read in the foreign language. It makes students depend on context clues and sentence meaning. This kind of item is constructed by deleting a word from a sentence, for example:

She quickly _______ her lunch

a. Drank ***b.*** Ate c. Drove d. Slept

c. Multiple – Choice Paraphrase

This kind of test offers much of the same advantages that multiple choice completion tests do and the contexts are much easier to prepare. Understanding is checked from the students by choosing the best synonym or paraphrase of the vocabulary item. For example:

They told us about the **savory** meal that they had eaten.

a. Broken ***b.*** Tasty c. Unhappy d. Helpless

d. Simple Completion (Words)

Words formation items require students to fill in missing parts of words that appear in sentences. These missing parts are usually prefixes and suffixes. For example:

When you write your check, make it pay_____ to my sister

The answer is payable.

6. Testing Pronunciation

Heaton (1990: 56) includes pronunciation into testing speaking skill. There are at least three techniques that can be used in testing pronunciation:

a. Pronouncing word in isolation

The important of listening in almost all test of speaking especially the pronunciation should never be underestimated.

b. Pronouncing words in a sentence

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

c. 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.

7. Testing Grammar

The specification of grammar test should be in line with the teaching syllabus if the syllabus lists the grammatical structure to be taught. When there is no such list, it becomes necessary to infer from textbook or other teaching materials. Heaton (1988: 34-50) suggested the techniques of testing grammar as follows:

a. Multiple Choice Items

This type of testing is favored by many constructors of grammar tests with incomplete statement type and a choice of four or five options.

b. Changing Words

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

c. Constructing Pairing and Matching Item

This type of item is usually consists of a short conversation then the students should match between the question and answer.

D. Types of Assessment

In this sub chapter the writerwill explain about type and form of assessment and testing. There are two types of assessment, informal and formal assessment (Brown,2000:384). Informal assessment can take a number of forms starting from incidental, unplanned comments and responses, along with coaching and other impromptu feedback to the student (Brown, 2000: 402). In this type of assessment, teachers recordstudents' achievement by some techniques that are not systematically made. Teachers can memorize what students do in the classroom based on their learning activities. Whereas, formal assessment are exercises or procedures specifically designed to tap into a store house of skills and knowledge and the purpose is to measure the students' language competence (Brown, 2000:402). Different from informal assessment, the formal assessment is intentionally made by teacher to get students' score to know their achievement and their progress. This assessment is done by teachers through making standard and official based on the rule.

In addition, Brown (2000:384) stated that there are two kinds of assessment based on the purpose and the times when the assessment is conducted; formative and summative assessment. Formative assessment intends to evaluate students in the process of forming their competencies and skills with the goal of helping them to continue that growth process (Brown, 2000:402). This assessment is conducted or done during teaching and learning process in the classroom, so there won't be a special time to conduct this kind of assessment because students' activities and responds during teaching and learning process will be used as the formative assessment. It purposes toknow students' product and progress of the teaching and learning process directly because it is conducted in every meeting, but this kind of assessment more emphasizes to know the students' progress rather than their product or achievement.

In addition, an assessment can be considered to be formative when teachers use it to check on the progress of theirstudents, to see how they have mastered what they should have learned, and then use this information to modify their future tea ching plans.

Summative assessment, then, aims to measure, or summarize, what students have grasped, and typically occurs at the end of a course or unit of instruction (Brown, 2000: 402-403).This kind of assessment is used by the teachers to measure and evaluate what students achieved during the process of teaching and learning in classroom, so it is conducted at the end of the semester or course.In short, formative assessment is done during the process of teaching and learning or in the middle of the semester,while summative is done in the end of the semester.

In doing summative test, teachers can use either multiple-choice test, short-answer test or even essay test. Each of them have different characteristics to be applied in evaluation, therefore, the detailed explanation of those kinds of test form is presented as follows:

1. Multiple-Choice Test

Multiple-choice Question test is the simplest test technique commonly used by test-makers. It can be used in any condition and situation, in any levelor degree of education. Actually, its simplicity relies on its scoring and answering because the examinees only need to choose one correct answer from the possible answers provided and the scorer only need to give one score for the correct answer and zero for the wrong answer.

According to Haladyna and Downing (1989b) in Osterlind (2002:164) that the use of multiple-choice formas generally leads to more content valid test score and interpretation. Yet,designing multiple-choice question is more complicated than essay items. Multiple-choice items may appear to be the simplest kind of item to construct but extremely difficult to design correctly. Multiple-choice item stake many forms, but their basic structure is that it has stems or the question itself, and a number of options-one which is correct, the others being distractors (Hughes, 2005:75).

In another case, Hughes (2005:76-78) states number of weaknesses of multiple-choice test that multiple-choice question is only recognition of knowledge. They make test takers can only guess to come with correct answer, and cheat easily. The technique severely restricts what can betested. It is very difficult to write successful items and the answer is restricted by the optional answer. In this case, test-takers can not elaborate their answer and understanding of the material because the answer is limited only by an optional answer.

Multiple-choice comes to be the first part of test packs faced by test-takers. When we want to analyze this item we can use statistical analysis asstated in the next chapter. Since there is only one right answer, the score canvery rapidly mark an item as correct and incorrect (Valette, 1967:6). Thus, wecan use simple codes to present the answer of test-takers. Score 1 presents correct answer chosen by students, and 0 presents wrong answer. If students choose a correct answer, we can note it by 1. And vice versa, if test-takers answer with wrong answer we note it with number 0.

2. Short-Answer Test

After test-takers have already answered the multiple choice items in first chapter of test-packs, in the next chapter they have to answer on short-answer items. The question is just the same, but in these items students are not given an optional answer. The answers are usually only one or two words. Those answers should be exactly correct, but the exactly correctans wer usually occursinonly listening and reading tests (Hughes,1989:79).

Regarding that English first semester test contains reading and writing skills, student's answer of this items especially on reading skill should exactly correct.Short-answer items deal with measurement of students' knowledge acquisition nd comprehension. It has two choices or formats, free and fixed.

Basically, there are two basic free formats. They are unstructured format and fill-in or completion format. Fixed choice format, then, consists of true-false,other two-choice, multiple-choice and matching (Tuckman, 1975:77). Short-answer items in English final semester test-packs used in this study are the items in which students should answer by writing down the answer in ashort and brief sentence. They are different from essay-test items.

In essay-test items, students should explore and elaborate their answer. For example, ifthe question is about structure and grammar, usually students should fill inthe blank with a complete sentence. Yet, in short-answer items what students should answer are usually not more than two or three words. As Valette(1967:8) states that this item may require one-word answer, such as brief responses to questions, or the filling in of missing elements.

In addition, in the short-answer items, the true answer has been determined byteachers so that students can not elaborate their answer. Both free choice and fixed choice items have previously determined correct response. In this formats, basically, measurement involves asking students a question that requires them to state or name the specific information or knowledge (Tuckman, 1975:77). Sometimes, the short-answer items are in the form of unstructured and completion/ fill-in format. In unstructured format, students can answer by a word, phrase or number. While in completion or fill-in format, students must construct their own response rather than choose an optional answer.

In order to assure to the objective nature of short-answer items, teacher must prepare a scoring system in advance (Valette, 1967:8). Teacher should give credit score to students' answer for misspelling of the world given. But since in short answer usually the answer is only one word, we can use the credit point the same as multiple choice. We can use the score 1 to presents students chosen correct answer and number 0 that presents incorrect answer. We only have to mark as 1 and 0 because the answer has been determined by test-maker and there is no optional answer for test-takers.

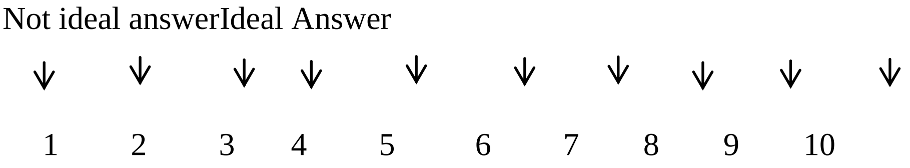
3. Essay Test

In English final test of elementary school, beside multiple choice andshort-answer items, there is one more test technique that is served to the test-takers in final semester. It is essay test. Different from short-answer items, essay test needs longer sentence to answer it. While short answer is the continuity of multiple choice items, essay-test items involve deep thinking about test-takers knowledge and understanding on material.

In language testing, it may include in students understanding on language structure and culture. It is supported by what Tuckman (1975:111) stated that “Essay items provide test-takers with the opportunity to structure and compose their own responses within relatively broad limits enable them to demonstrate their ability to apply knowledge and to analyze, to synthesize, and to evaluate new information in the light of their knowledge.”

The scoring system of this item will be very different from scoring objectives items or multiple-choice. In objective items, the score of each number is exact and all the same from number to number. Whereas, in essay items, what we should do, first, is determining the ideal answer even though no correct and wrong answer at all. The ideal answer then should be scored asthe highest score. The far answers of students go beyond it will be the lowest score it is. Teachers then should create interval scale to score the highest and the lowest one on each item. Interval scale will be going like picture below:

Diagram 2.1: Interval Scale for Essay-test items Scoring



The interval scale then can be used to measure how far students understand the material. If the students get higher score, it means that they understand more on the material. Teachers have an authority to determine interval scale number between ideal and not-ideal answer. It can be a scale from 0 until 10 like the scale above, or 0 until 3 or 5 based on their preferences. It may be decided by calculating every score of every item, from the multiple-choice questions, short-answer items, and the last one is essay-test items.

E. Item Analysis

Item Analysis is related to the several items of statistical analysis in analyzing characteristics and features of a test. They consist of validity, reliability, level of difficulty, discriminating power, and distribution of distractors.

a. Validity

Validity is the extent to which the test actually measures what it is intended to measure (Brown, 2000:387) it is also the extent to which inferences made from assessment results are meaningful, and useful in term of the purpose of the assessment.

Validity can also be defined as the extent to which the instrument measures what it should be measured, so the test should test what the writer or teacher wants to test the students. The expert should look into whether the test content is representative of the skills that are supposed to be measured. This involves looking into the consistency between the syllabus content, the test objective and the test contents. If the test contents cover the test objectives, which in turn are representatives of the syllabus, it could be said that the test possesses content validity (Brown, 2002: 23-24).

Brown’s idea is supported by Hughes (2005:26), who stated that a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc in which it is meant to be concerned. It means that a test will have content validity.

b. Reliability

Reliability refers to the consistency of score. A reliable test is that a test which has consistency of score, it means that the test can produce similar score if it is conducted for the second time or more times to the same students at different time. Bachman (2004:153) states that reliability is consistency of measures a cross different conditions in the measurement procedures. So that, the more similar the scores are, the more reliable the test is.

In line with the reliable test, in order to quantify the reliability of the test items, the teachers may quantify it in the form of a reliability coefficient. The ideal reliability coefficient is that ±1. It means that if the test has a reliability of coefficient closes to ±1, so the test can produce almost the same result of a particular set of test-takers regardless when it happened to be administered. On the other hand, a test which has a reliability of coefficient of zero would give sets of result quite unconnected with each other.

c. Item Facility Analysis

A good test is a test which is not too easy nor too difficult. Thus, the test should be standard and fulfill the characteristics of a good test. The number that shows the level difficulty of a test can be said as difficulty index (Arikunto, 2012:222). In this index there are minimum and maximum scores. The lower index of a test, the more difficult the test is. And vice versa, the higher the test, the easier it is.

The categorizing of index of difficulty is proposed by Arikunto (2012:225) that a test items is called to be difficult if the number of P (index of difficult) is between 0.00-0.300. A test item is in range of sufficient or fair if the index of difficulty is between 0.31-0.70. Then, it is called as easy test if the index is between 0.71-1.00.

There are some factors that every test constructors must consider in constructing difficulty level of test items. Mehren and Lehmen (1984) point out that the concept of difficulty or the decision of how difficult the test should depends on variety of factors, notably 1) the purpose of the test, 2) students ability level , and 3) the age of grade.

d. Item Discrimination Analysis

It is the extent to which an item differentiates between high and low-ability test-takers. Discrimination is important because if the test-items can discriminate more, they will be more reliable. It can also be defined as the ability of a test to separate master students and non-master students (Arikunto, 2012:226). A master student is a student with higher scores of test, and a non-master student is a student with lower scores on the test given.

The same as the term of difficulty level, discrimination has discrimination index. It is an indicator of how well an item discriminates between weak candidates and strong candidates (Hughes, 2005:226). This index is used to measure to the ability of a test in discriminating the upper and lower group of students. Upper students are students who answer with correct answer, and lower group are students with wrong answer.

Different from difficulty index, the negative index of discrimination power shows that the questions identify high group students as poor students and low group students as smart students. Whereas, a good question is actually a question that can be answered by upper group and cannot be answered correctly by the lower group.

In line with the level of difficulty of the test items and its relation to the item discriminators is that an item will have poor index difficulty if it cannot differentiate between smart students and poor students. It happens if smart students and poor students have the same score on the same item or even the poor students have higher score than the upper students. Conversely, an item that garners correct responses from most the high-ability group and incorrect responses from most of the low ability group has good discrimination power (Brown, 2004:59).

e. Distractor Efficiency Analysis

In addition to calculating discrimination indices and facility values, it is necessary to analyze the performance of distractors (Hughes, 2005:228). It is defined as the distribution of testee in choosing the optional answer (distractors) in multiple choice questions (Arikunto, 2012:233). It can be obtained by calculating the number of testee in choosing the distractors. We can calculate this form by seeing the answer form done by students. The distractors are good if chosen by minimum 5% of the number of test takers.

One way to study responses to distractors is with frequency table that tells us the proportion of students who selected a given distractor. Distractors that are not chosen by any examinees should be replaced or removed. Distractors that do not work for example are chosen by very few test-takers should be replacing by better ones, or the item should be otherwise modified or dropped (Hughes, 2005:228). They should be discarded because they are chosen by very few test-takers from both groups.It means that they cannot function properly.

CHAPETR III

RESEARCH METHOD

This chapter consists of six sub chapters. They are research design, population and sample, time and place of the study research instrument, data collecting method, and data analysis.

A. Research Design

Research design is defined as the strategy or the way how the researcher gets valid data, analyze them, and finally come to the answer of the research problem. In this research, the research design used was descriptivewith quantitative data. It was descriptive since the aim of this study is to present and describe the quality of the English final test by analyzing the test items with quantitative approach because the data deals with score and number and the result of this study was generalisable since there was a sample analysis.

B. Time and Place of The Study

The researcher conducted a research on February 17, 2015 in SMAN 1 Kedungwaru. The researcher asked the documentation of English final test at the first semester of 12th grade students, students' answer sheets, answer key, and the syllabus to the English teacher of 12th grage students of SMAN 1 Kedungwaru.

C. Population and Sample

Population is the group to which the 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". The population of this study was English final test of the first semester students grade XII of SMAN 1 Kedungwaru which consists of two kinds of test-packs; test-pack A and B; each of them consist of 40 items; and the answer sheets of the students did the test.

Sample is small group which is taken from the population and it is observed. According to Ary *et al* (2002: 138) "Sample is a part of population, which wants to be analyzed". In this study, the samples were the population itself, English Final test of the first semester students grade XII of SMAN 1 Kedungwaru because the population in this research was 40 items for each test-pack A and B. For the students' answer sheet, the researcher took 40 students from each test-pack who did the test as the sample, thus there were 80 answer sheets of the students; 40 answer sheets of test-pack A and 40 answer sheets of test-pack B. The researcher took 40 students as the participant because this kind of research needs at least 30 participants per group to be analyzed as the sample and the sample was taken through random sampling so that the result of the sample analysis can be used for generalization.

D. Research Instrument

The term instrument in a study refers to any kind of tools used by the researcher to get the information or data. Fraenkel (2005:112) states "Instrument is the device the researcher uses to collect data". The instrument in this study was document in the form of english final test, students' answer sheets, key answer and syllabus. Lincoln and Guba (1985:57) defined a document as "any written or recorded material" not prepared for the purposes of the evaluation or at the request of the inquirer.

E. Data Collecting Method

The method of collecting data can be considered as the way reseracher get the data. In this study, the researcher used documentation as the method of collecting data since the data were in the form of document. Tanzeh (2011:93) states "Documentation is collecting data by looking or writing a report that available such as written material or film".

F. Data Analysis

In this study, the data was analyzed quantitatively. The quantitative data analysis was done by analyzing the test items and students' answer sheets. There are five points of item analysis; validity, reliability, item facility, discrimination power and distractor efficiency.

1. Validity

Measuring the validity of a test is not as easy as measuring the reliability, item facility, discrimination power, and distractor efficiency because the validity cannot be measured using formula. In order to ensure that a test is valid, the researcher measured two types of validation; content validity and construct validity.

A test can be claimed as a valid test in term of its content if the test items can measure all the materials have been taught, not other or outside the given material. Hughes (1989: 22) stated that a test is said to have content validity if its content constitutes a representative sample of language skills, structures, etc. with which it is meant to be concerned.

To know whether the test items have good content validity or not, the researcher used the syllabus to get the clear specification of the skills or components or materials that it is meant to cover, then compared the test content andthe specification stated in the syllabus. At last, the researcher gave the percentage of skills being tested based on the specification provided.

Besides the content validity of the test, it is also necessary to know the construct validity of the test items in language testing. A test can be said to have construct validity if the test is created based on the underlying ability of each skill and component being tested. Hughes (1989: 26) added that a test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure because the word “construct” refers to any underlying ability which is hypothesized in a theory of language ability, so the researcher used the language testing theory of language ability to know whether the test has good construct validity or not.

In this study, the researcher analyzed the testing techniques used in a test, then connected it to the language testing theory to know whether the testing techniques used in the test are already appropriate to the language testing theory or not. For instance, writing test shown in the test items number 2, 6, 8, and 9, provided the students to complete the blank paragraph with the correct vocabularies provided, speaking test shown in the test items number 3, 4, 15, 27 asked the students to give correct response to a certain dialogue, reading test shown in the test items number 23, 12, 14, 13, was about deciding the similar meaning of the certain words. Listening test was not shown in the test item, it means that the test items were less of construct validity for listening skill, whereas listening skill was also mentioned in the syllabus material that it must also be achieved.

2. Reliability

In order to measure the reliability of the test items, the researcher used the KR-20 formula because this formula requires test administration only once and the scoring is one correct answer is given point 1, while incorrect answer is given 0, thus this formula is appropriate for calculating the reliability of multiple choice test form. In addition, Fraenkel and Wallen (2008:156) stated that KR-20 doesn't require the assumption that all items are of equal difficulty .

- KR-20 Formula

$$\boxed{r_{11} = \left[\frac{n}{n-1}\right]\left[\frac{s_t^2 - \sum p_1 q_1}{s_t^2}\right]}$$

Where:

r_{11} = reliability coefficient

n = number of test items

s_t^2 = standard deviation

p_1 = proportion of the right respond

q_1 = proportion of the wrong respond

After calculating the reliability of the test items, the researcher classified the reliability coefficient which taken from Sudijiono (1996: 209-230), as the table follows:

Table 3.1 Classification of Reliability Test

Reliability Test Coefficient	Classification
0.99-1.00	More highly
0.70-0.89	High
0.50-0.69	Fair
0.30-0.49	Low
<0.30	Very low

3. Measuring the Item Facility

To measure the item facility of level of difficulty of the test items, the researcher used the following formulas:

$$P = \frac{B}{JS}$$ (Arikunto, 2012: 223)

Where:

P = Item Facility (Level of difficulty)

B = Number of test-takers answering the item correctly

JS = number of test-takers responding to that item

To know the classification of the difficulty level, the researcher used the classification referred by Arikunto (2012:225). Here is the following classification and interpretation of difficulty level:

Table 3.2 Classification of Difficulty Indices

Difficulty Level	Classification
0.00-0.30	Difficult
0.31-0.70	Fair
0.71-1.00	Easy

4. Measuring Discrimination Power

In order to measure the discrimination power of each item, the researcher needed to separate the students into upper and lower group in order to be applied in the following formula:

$$\boxed{DP = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B}$$

(Arikunto, 2012:228)

Where:

DP= Discrimination Power

J =Number of Test-takers

J_A =Totalparticipant of top test-takers

J_B =Total participant of bottom test-takers

B_A= Number of top test takers that have correct answer

B_B =Number of bottom test takers that have correct answer

$P_A = \frac{B_A}{J_A}$ = Proportion of the number of top class answering correctly

$P_B = \dfrac{B_B}{J_B}$ = Proportion of bottom class answering correctly

According to Arikunto (2012:232), here is the classification and interpretation of discrimination index:

Table 3.3 Classification and Interpretation of Discrimination Indices

Discrimination Index	Classification
0.71-1.00	Excellent
0.41-0.70	Good
0.21-0.40	Satisfactory
$\leq$ 0.20	Poor
Negative value on D	Very Poor

5. Measuring Distractor Efficiency

The distribution of distractors means the distribution of alternative answers. The importance of calculating it is to know how well the distractors work in distracting the students’ answer. A good distractor is that it has the distribution index of more than 5% of the total examinees number. Arikunto (2012: 238) points out that a distractor can be said to have functioned well when it is chosen by at least 5% of the total examinees. If the index of this is 0, thus the distractor should be discarded or eliminated.

CHAPTER IV

RESEARCH FINDING AND DISCUSSION

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

A. The Description of Data

1. Validity

In this research, the researcher used two types of validity; content validity and construct validity.

a. Content Validity

The researcher analyzed the content validity of the first semester english final test of the 12th grade students of SMAN 1 Kedungwaru in academic year 2014/2015. It has been known that a good test items must have content validity and content validity itself must be upon on careful analysis of the outline of the course. Furthermore, it is expected that the test items must represent each proportion of the material stated in the outline of the course adequately. In addition, content validity analysis deals with the comparison of what was tested by the test and what actually to be tested. To know how good the content validity of the first semester English final test of the 12th grade students of SMAN 1 Kedungwaru was, the researcher compared the syllabus content to each test items as table 4.1 and 4.2:

Table 4.1 The Appropriateness of The First Semester English Final Test with The English Syllabus of SMAN 1 Kedungwaru

Skill	The Basic Competences in Syllabus	Number of Item	
		Test A	Test B
Listening	**1.** Merespons makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterimadalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: **mengakui kesalahan, berjaniji, menyalahkan, menuduh,** mengungkapkan keingintahuan dan hasrat, dan menyatakan berbagai sikap.		
	2. Merespons makna dalam teks fungsional pendek: **pengumuman (announcement)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
	3. Merespons makna dalam teks monolog yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: **narratives, explanation, dan discussion.**		
	4. Merespons makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterima dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: **mengusulkan, memohon, mengeluh,** membahas kemungkinan atau untuk melakukan sesuatu dan **memerintah.**		
	5. Merespons makna dalam teks fungsional pendek: **pesan telepon (telephone message)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
	6. Merespons makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterimadalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: mengusulkan, memohon, mengeluh, **membahas kemungkinan atau untuk melakukan sesuatu** dan memerintah, serta **mengungkapkan keingintahuan dan hasrat, dan menyatakan berbagai sikap.**		
	7. Merespons makna dalam teks fungsional pendek: **iklan layanan masyarakat (public service announcement)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
Speaking	1. Mengungkapkan makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterimadalam	8	7, 14

	konteks kehidupan sehari-hari dan melibatkan tindak tutur: **mengakui kesalahan, berjaniji, menyalahkan, menuduh,** mengungkapkan keingintahuan dan hasrat, dan menyatakan berbagai sikap.		
	2. Mengungkapkan makna dalam teks monolog yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: **narratives, explanation, dan discussion.**		
	3. Mengungkapkan makna dalam teks fungsional pendek: **pengumuman (announcement)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
	4. Mengungkapkan makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterima dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: **mengusulkan, memohon, mengeluh,** membahas kemungkinan atau untuk melakukan sesuatu dan **memerintah.**	7	
	5. Mengungkapkan makna dalam teks fungsional pendek: **pesan telepon (telephone message)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
	6. Mengungkapkan makna dalam percakapan transaksional (to get things done) dan interpersonal (bersosialisasi) resmi dan berlanjut (sustained) secara akurat, lancer dan berterima dalam konteks kehidupan sehari-hari dan melibatkan tindak tutur: mengusulkan, memohon, mengeluh, **membahas kemungkinan atau untuk melakukan sesuatu** dan memerintah, serta **mengungkapkan keingintahuan dan hasrat, dan menyatakan berbagai sikap.**		
	7. Mengungkapkan makna dalam teks fungsional pendek: **iklan layanan masyarakat (public service announcement)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
Reading	1. Merespons makna dan langkah retorika dalam esei yang menggunakan ragam bahasa tulis secara akurat, lancer dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: **narratives (narrative text, modal perfect, conditional using 'wish', kosakata yang terkait dengan topik yang dipelajari) explanation (explanation text, passive voice, kosakata yang terkaitdengan topic yang dipelajari) , dan discussion (discussion texs, contrastive conjunction, dankosakata yang terkaittopik yang dipilih).**	3,4,5,6,9 101114, 15,18,19 20,21,22 23,24,25 26,31,32 33,34, 35	3,4,5,6,9 10,1112, 13,15,16 1718,24, 25,26,27 28,29,30 31,32,35 36
	2. Merespons makna dalam teks fungsional pendek: **pengumuman (announcement)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari dan untuk mengakses ilmu pengetahuan.	1,2	1,2,19, 20,21, 22
	3. Merespons makna dalam teks fungsional pendek: **surat resmi**	16,17,27	

	(formal letter) misalnya iklan, undangan, dll resmi dan tidak resmi yang menggunakan ragam bahasa lisan secaraakurat, lancar, dan berterima dalam konteks kehidupan sehari-hari dan untuk mengakses ilmu pengetahuan.	28,29,30	
	4. Merespons makna dalam teks fungsional pendek: **leaflet (misalnya banner, poster, pamphlet, dll)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secaraakurat, lancar, dan berterima dalam konteks kehidupan sehari-hari dan untuk mengakses ilmu pengetahuan.		
Writing	1. Mengungkapkan makna dan langkah retorika dalam esei yang menggunakan ragam bahasa tulis secara akurat, lancer dan berterima dalam konteks kehidupan sehari-hari dalam teks berbentuk: **narratives (narrative text, modal perfect, conditional using 'wish') explanation (explanation text), dan discussion (discussion texs, contrastive conjunction).**	12,13,36 37,38,39 40	8,23,33, 34,37,38 39, 40
	2. Mengungkapkan makna dalam teks fungsional pendek: **surat resmi/ formal letter, misalnya pengumuman, iklan, undangan dll** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		
	3. Mengungkapkan makna dalam teks fungsional pendek: **surat resmi/ formal letter (misalnya banner, poster, pamphlet, dll)** resmi dan tidak resmi yang menggunakan ragam bahasa lisan secara akurat, lancar, dan berterima dalam konteks kehidupan sehari-hari.		

Based on the table above, it can be seen that the test items of test-package A did not cover all material in the syllabus such as the the second, third, fifth, sixth and seventh material in speaking; the fourth material in reading; and the second and third material in writing. Moreover, no material in listening was included in the test items. It was also happened to the test items of test-package B which did not cover all material in the syllabus as well. The second, third, fourth, fifth, sixth, and seventh material in speaking and the third and fourth material in reading and writing were not included in the test items.

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

Table 4.2.The Percentage of Skill Tested in English Final Testof SMAN 1 Kedungwaru

Language Skill	The Percentage of Skill Being Tested	
	Test-Package A	Test-Package B
Listening	0%	0%
Speaking	2/40 x 100 % = 5%	2/40 x 100% = 5%
Reading	31/40 x 100% = 77.5%	30/40 x 100% = 80%
Writing	7/40 x 100 % = 17.5 %	8/40 x 100 % = 20 %

b. Construct Validity

The second analysis was construct validity. Hughes (1989: 26) stated that a test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure because the word “construct” refers to any underlying ability which is hypothesized in a theory of language ability, so the researcher used the language testing theory of language ability 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 Technique Used in English Final Testof SMAN 1 Kedungwaru

Test-Package A	**Test-Package B**
Speaking Test The speaking test was shown in numbers 7 and 8 • Item numbers 7 and 8 used the blank dialogue and asked students to response the dialogue/ expression.	**Speaking Test** The speaking test was shown in numbers 7 and 14 • Item number 7 used the blank dialogue and asked students to response the dialogue • Test item number 14 used the dialogue and asked the students to categorize the dialogue
Reading Test The reading test was shown in numbers 1,2, 3,4,5,6, 9,10,11, 14,15,18, 19,20,21, 22,23,24,25,26,27, 28, 29, 30, 31 32,33, 34, and 35 • Item numbers 1,2, 16, 27, 28, 29, and 30 asked the students to choose the correct answer related to the information of announcement and letter • Item number 4 asked the students to identify the meaning of the underlined sentence in the text. • Item numbers 22, 26, and 34 asked the students to guess the meaning of the unfamiliar word • Item numbers 3, 5, 6, 9, 10, 11, 14, 15, 17, 18, 19, 20, 21, 23, 24, 25, 31, 32, 33 and 35 asked the students to identify the topic, purpose, the generic structure and moral value of the text.	**Reading Test** The reading test was shown in numbers 1,2, 3,4,5,6, 9,10,11, 12, 13, 15,16, 17, 18,19,20, 21,22,24,25,26,27, 28, 29, 30, 31 32,35, and 36 • Item numbers 1, 2, 19, 20, 21, and 22 asked the students to choose the correct answer related to the information of announcement • Item numbers 3, 5, 6, 9, 10, 11, 12, 15, 16, 17, 18, 24, 25, 26, 27, 28, 29, 30, 32, 35, 36, asked the students to identify the topic, purpose, the generic structure and the value of the text; narrative, explanation and discussion text. • Item numbers 4, 13, and 31 asked the students to guess the meaning of the unfamiliar word.
Writing Test The writing test was shown in numbers 12,13, 36, 37, 38, 39, and 40 • Item numbers 12, and 13 asked students to complete the sentence about conditional sentence with the correct phrase or clause • Item numbers 36, 37, 38, 39 and 40 asked students to complete the blank paragraph with the vocabularies provided	**Writing Test** The writing test was shown in numbers 8, 23, 33, 34, 37, 38, 39, and 40 • Item numbers 8, and 23 asked students to give the meaning on a sentence using conditional “wish”. • Item number 7 asked students to complete the sentence about conditional sentence with the correct phrase or clause. • Item numbers 33, 34, 38, 39 and 40 asked students to complete the blank paragraph with the vocabularies provided

The technique of overall English skill test was multiple-choice test. The researcher found that the speaking test was dominated by the questions about responding and categorizing a dialogue. In testing reading, the researcher found that the test provided the students to choose the correct answer about any information related to the text, the purpose of the text, moral value, identify the topic and generic structure of the text. Then, the students also had to choose the right answer for the meaning or similar meaning of the unfamiliar words. The last was testing writing. In testing writing, the researcher found that the test provided the students to choose the appropriate vocabularies to complete the blank paragraph and some about grammar.

2. Reliability

The next was the reliability analysis. Reliability refers to the stability of the score. The reliability can be estimated by formula Kuder Richardson (KR 20):

$$\boxed{r_{11} = \left[\frac{n}{n-1}\right]\left[\frac{s_t^2 - \sum p_1 q_1}{s_t^2}\right]}$$

Where: r_{11}=reliability coefficient

n = number of test items

s_t^2 = standard deviation

p_1 = proportion ofthe right respond

q_1 = proportion of the wrong respond

Before computing the reliability, the standard deviation must be computed first by using the following formula:

$$S = \sqrt{\frac{\Sigma(X-\mu)2}{N}}$$

Where:

S= Standard deviation

X= Individual score

µ= Population mean

N= Number of the students

Table 4.4 The preparatory to compute the standard deviation of Test A

No.	**Name**	**X**	**µ**	**(X- µ)**	**(X- µ)^2**
1.	BPH	**90**	83.5	6.5	42.25
2.	AF	**90**	83.5	6.5	42.25
3.	ABP	**85**	83.5	1.5	2.25
4.	AG	**87.5**	83.5	4	16
5.	ARN	**92.5**	83.5	9	81
6.	ERS	**85**	83.5	1.5	2.25
7.	ADB	**72.5**	83.5	-11	121
8.	HMN	**90**	83.5	6.5	42.25
9.	ACK	**90**	83.5	6.5	42.25
10.	DS	**90**	83.5	6.5	42.25
11.	IPY	**87.5**	83.5	4	16
12.	HP	**90**	83.5	6.5	42.25
13.	NDO	**90**	83.5	6.5	42.25
14.	VVDP	**90**	83.5	6.5	42.25
15.	RM	**77.5**	83.5	-6	36
16.	SW	**92.5**	83.5	9	81
17.	NP	**92.5**	83.5	9	81
18.	NF	**92.5**	83.5	9	81
19.	WF	**77.5**	83.5	-6	36
20.	SDA	**90**	83.5	6.5	42.25
21.	MUAA	**92.5**	83.5	9	81
22.	UAS	**92.5**	83.5	9	81
23.	VBDP	**90**	83.5	6.5	42.25
24.	SCD	**85**	83.5	1.5	2.25
25.	TD	**90**	83.5	6.5	42.25
26.	DALP	**90**	83.5	6.5	42.25

27.	BAN	**87.5**	83.5	4	16
28.	MHH	**70**	83.5	-13.5	182.25
29.	PBM	**82.5**	83.5	-1	1
30.	NANH	**87.5**	83.5	4	16
31.	AN	**75**	83.5	-8.5	72.25
32.	MAE	**72.5**	83.5	-11	121
33.	JBP	**75**	83.5	-8.5	72.25
34.	AGW	**67.5**	83.5	-16	256
35.	BAR	**72.5**	83.5	-11	121
36.	KYN	**67.5**	83.5	-16	256
37.	MFHP	**72.5**	83.5	-11	121
38.	SEP	**72.5**	83.5	-11	121
39.	USW	**72.5**	83.5	-11	121
40.	NA	**72.5**	83.5	-11	121
		∑**X**=3340	83.5		∑**(X-µ)**^{2}= 2822.50

Therefore, the standard deviation is

$$S = \sqrt{\frac{\Sigma(X-\mu)2}{N}}$$

$$= \sqrt{\frac{2822.5}{40}}$$

$$= 8.4$$

After finding the rsult of standard deviation, the reliability can be computed using KR-20 formula.

Table 4.5 The Table to Compute The Reliability by Using KR-20 Formula

Item	**Np**	**P_1**	**Nq**	**Q_1**	**P_1 Q_1**
1	0	0	40	1	0
2	40	1	0	0	0
3	39	0.975	1	0.025	0.02438
4	40	1	0	0	0
5	29	0.725	11	0.275	0.199
6	40	1	0	0	0
7	13	0.325	27	0.675	0.21938
8	38	0.95	2	0.05	0.0475
9	40	1	0	0	0
10	39	0.975	1	0.025	0.02438
11	38	0.95	2	0.05	0.0475

12	0	0	40	1	0
13	27	0.675	13	0.325	0.21938
14	40	1	0	0	0
15	25	0.625	15	0.375	0.2338
16	39	0.975	1	0.025	0.02438
17	22	0.55	18	0.45	0.2475
18	40	1	0	0	0
19	27	0.675	13	0.325	0.21938
20	36	0.9	4	0.1	0.09
21	40	1	0	0	0
22	40	1	0	0	0
23	40	1	0	0	0
24	40	1	0	0	0
25	39	0.975	1	0.025	0.02438
26	40	1	0	0	0
27	40	1	0	0	0
28	40	1	0	0	0
29	40	1	0	0	0
30	40	1	0	0	0
31	37	0.925	3	0.075	0.0638
32	35	0.875	5	0.125	0.10938
33	40	1	0	0	0
34	40	1	0	0	0
35	40	1	0	0	0
36	40	1	0	0	0
37	1	0.025	39	0.975	0.02438
38	27	0.675	13	0.325	0.21938
39	28	0.7	12	0.3	0.21
40	27	0.675	13	0.325	0.21938
					$\sum p_1q_{1=}$ 2.46728

Therefore, the reliability is:

$$r_{11} = \left[\frac{n}{n-1}\right]\left[\frac{s_t^2 - \sum p_1 q_1}{s_t^2}\right]$$

$$r_{11} = \left[\frac{40}{40-1}\right]\left[\frac{8.4-2.4678}{8.4}\right]$$

$$r_{11} = \left[\frac{40}{39}\right]\left[\frac{5.9322}{8.4}\right]$$

r_{11}= [1.02564] [0.70621]

r_{11}= 0.72432

In order to strengthen the result of the reliability coefficient of test-package A after computed manually, the researcher also used SPSS aplication to compute the reliability coefficient of test-package A, and the result showed that the reliability coefficient computed manually was almost equal with the reliability coefficient computed by SPSS that is 0.72434. It means that the reliability of test-package A is fair.

Table 4.6 The preparatory to compute the standard deviation ofTest B

No.	**Name**	**X**	**µ**	**(X- µ)**	**(X- µ)^2**
1.	DFA	**82.5**	77.875	4.625	21.390625
2.	MRS	**82.5**	77.875	4.625	21.390625
3.	RFL	**75**	77.875	-2.875	8.265625
4.	AM	**75**	77.875	-2.875	8.265625
5.	RK	**60**	77.875	-17.875	319.515625
6.	BS	**57.5**	77.875	-20.375	415.140625
7.	AFD	**82.5**	77.875	4.625	21.390625
8.	RF	**82.5**	77.875	4.625	21.390625
9.	WFR	**77.5**	77.875	-0.375	0.140625
10.	NH	**67.5**	77.875	-10.375	107.640625
11.	YO	**75**	77.875	-2.875	8.265625
12.	LBS	**87.5**	77.875	9.625	92.640625
13.	IF	**77.5**	77.875	-0.375	0.140625
14.	BAK	**80**	77.875	2.125	4.515625
15.	PD	**77.5**	77.875	-0.375	0.140625
16.	SP	**77.5**	77.875	-0.375	0.140625
17.	SPA	**80**	77.875	2.125	4.515625
18.	IW	**87.5**	77.875	9.625	92.640625
19.	KT	**85**	77.875	7.125	50.765625
20.	EFS	**67.5**	77.875	-10.375	107.640625
21.	FT	**72.5**	77.875	-5.375	28.890625
22.	EGW	**70**	77.875	-7.875	62.015625
23.	GAN	**77.5**	77.875	-0.375	0.140625
24.	CA	**70**	77.875	-7.875	62.015625
25.	DY	**72.5**	77.875	-5.375	28.890625
26.	MAR	**82.5**	77.875	4.625	21.390625
27.	MGA	**85**	77.875	7.125	50.765625

28.	DRS	**80**	77.875	2.125	4.515625
29.	GF	**75**	77.875	-2.875	8.265625
30.	AK	**85**	77.875	7.125	50.765625
31.	NHH	**75**	77.875	-2.875	8.265625
32.	NA	**85**	77.875	7.125	50.765625
33.	AYA	**82.5**	77.875	4.625	21.390625
34.	ACD.	**87.5**	77.875	9.625	92.640625
35.	JNA	**82.5**	77.875	4.625	21.390625
36.	FMAW	**87.5**	77.875	9.625	92.640625
37.	SRA	**82.5**	77.875	4.625	21.390625
38.	SDA	**85**	77.875	7.125	50.765625
39.	RYR	**62.5**	77.875	-15.375	236.390625
40.	ADA	**77.5**	77.875	-0.375	0.140625
		∑X= 3115	77.875		**∑(X- μ)**^{2}= 2219.38

Therefore, the standard deviation is

$$S = \sqrt{\frac{\Sigma(X-\mu)2}{N}}$$

$$= \sqrt{\frac{2219.38}{40}}$$

$$= 7.45$$

Table 4.7 The Table to Compute The Reliability By Using KR-20 Formula

Item	**Np**	**P_1**	**Nq**	**Q_1**	**P_1 Q_1**
1	40	1	0	0	0
2	4	0.1	36	0.9	0.09
3	40	1	0	0	0
4	39	0.975	1	0.025	0.02438
5	35	0.875	5	0.125	0.10938
6	29	0.725	11	0.275	0.19938
7	39	0.975	1	0.025	0.02438
8	26	0.65	14	0.35	0.2275
9	15	0.375	25	0.625	0.23438
10	38	0.95	2	0.05	0.0475
11	37	0.925	3	0.075	0.06938
12	17	0.425	23	0.575	0.24438
13	26	0.65	14	0.35	0.2275
14	40	1	0	0	0
15	40	1	0	0	0
16	38	0.95	2	0.05	0.0475

17	35	0.875	5	0.125	0.10938
18	40	1	0	0	0
19	40	1	0	0	0
20	38	0.95	2	0.05	0.0475
21	40	1	0	0	0
22	38	0.95	2	0.05	0.0475
23	25	0.625	15	0.375	0.2348
24	37	0.925	3	0.075	0.06938
25	40	1	0	0	0
26	33	0.825	7	0.175	0.14437
27	37	0.925	3	0.075	0.06938
28	37	0.925	3	0.075	0.06938
29	12	0.3	28	0.7	0.21
30	37	0.925	3	0.075	0.06938
31	40	1	0	0	0
32	39	0.975	1	0.025	0.02438
33	2	0.05	38	0.95	0.0475
34	34	0.85	6	0.15	0.1275
35	28	0.7	12	0.3	0.21
36	21	0.525	19	0.475	0.24938
37	2	0.05	38	0.95	0.0475
38	23	0.575	17	0.425	0.24438
39	24	0.6	16	0.4	0.24
40	40	1	0	0	0
					$\sum p_1q_1$= 3.93679

Therefore, the reliability is:

$$r_{11}=\left[\frac{n}{n-1}\right]\left[\frac{s_t^2-\sum p_1q_1}{s_t^2}\right]$$

$$r_{11}=\left[\frac{40}{40-1}\right]\left[\frac{7.45-3.93679}{7.45}\right]$$

$$r_{11}=\left[\frac{40}{39}\right]\left[\frac{3.51321}{7.45}\right]$$

r_{11}= 0.48366

In order to strengthen the result of the reliability coefficient of test-package B after computed manually, the researcher also used SPSS aplication to compute the reliability coefficient of test-package A, and the result showed that the reliability coefficient computed manually was almost equal with the reliability coefficient computed by SPSS that is 0.48376. It means that the reliability of test-package B is low.

3. Level of Difficulty

The level of difficulty shows how easy or difficult a test is. It can be seen through the number of the students can answer correctly and from which group; upper or lower students. The level of difficulty can be estimated by using the following formula:

$$P = \frac{B}{JS} \text{ (Arikunto, 2012: 223)}$$

Where:

P = Item Facility (Level of difficulty)

B = Number of test-takers answering the item correctly

JS = number of test-takers responding to that item

Arikunto (2012:225) stated the classification of the difficulty level of the test items as follows:

Difficulty Level	Classification
0.00-0.30	Difficult
0.31-0.70	Fair
0.71-1.00	Easy

Based on the classification and interpretation of difficulty level proposed by Arikunto, here is the classification and interpretation of difficulty level of english final test of the 12th grade students of SMAN 1 Kedungwaru:

Table 4.8 The Presentation of Level of Difficulty of Test-Package A

Item	B	JS	IF = B/JS	Classification
1	0	40	0	Difficult
2	40	40	1	Easy
3	39	40	0.975	Easy
4	40	40	1	Easy
5	29	40	0.725	Easy
6	40	40	1	Easy
7	23	40	0.575	Fair
8	38	40	0.95	Easy
9	0	40	0	Difficult
10	39	40	0.975	Easy
11	39	40	0.975	Easy
12	0	40	0	Difficult
13	27	40	0.675	Fair
14	40	40	1	Easy
15	25	40	0.625	Fair
16	39	40	0.975	Easy
17	32	40	0.8	Easy
18	40	40	1	Easy
19	27	40	0.675	Fair
20	36	40	0.9	Easy
21	40	40	1	Easy
22	40	40	1	Easy
23	39	40	0.975	Easy
24	40	40	1	Easy
25	40	40	1	Easy
26	40	40	1	Easy
27	40	40	1	Easy
28	40	40	1	Easy
29	40	40	1	Easy
30	40	40	1	Easy
31	36	40	0.9	Easy
32	37	40	0.925	Easy
33	40	40	1	Easy
34	40	40	1	Easy
35	40	40	1	Easy
36	40	40	1	Easy
37	1	40	0.025	Difficult
38	27	40	0.675	Fair
39	27	40	0.675	Fair
40	27	40	0.675	Fair

Table 4.9. The Presentation of Level of Difficulty of Test-Package B

Item	B	JS	P = NP/N	Classification
1	40	40	1	Easy
2	4	40	1	Easy
3	40	40	1	Easy
4	39	40	0.975	Easy
5	36	40	0.9	Easy
6	29	40	0.725	Easy
7	39	40	0.975	Easy
8	26	40	0.65	Fair
9	15	40	0.375	Fair
10	38	40	0.95	Easy
11	37	40	0.925	Easy
12	17	40	0.425	Fair
13	26	40	0.65	Fair
14	40	40	1	Easy
15	40	40	1	Easy
16	40	40	1	Easy
17	36	40	0.9	Easy
18	40	40	1	Easy
19	40	40	1	Easy
20	38	40	0.95	Easy
21	40	40	1	Easy
22	39	40	0.975	Easy
23	25	40	0.625	Fair
24	37	40	0.925	Easy
25	40	40	1	Easy
26	33	40	0.825	Easy
27	37	40	0.925	Easy
28	37	40	0.925	Easy
29	12	40	0.3	Fair
30	27	40	0.674	Fair
31				
32	39	40	0.975	Easy
33	2	40	0.05	Difficult
34	24	40	0.6	Fair
35	28	40	0.7	Fair
36	21	40	0.525	Fair
37	3	40	0.075	Difficult
38	23	40	0.575	Fair
39	1	40	0.025	Difficult
40	40	40	1	Easy

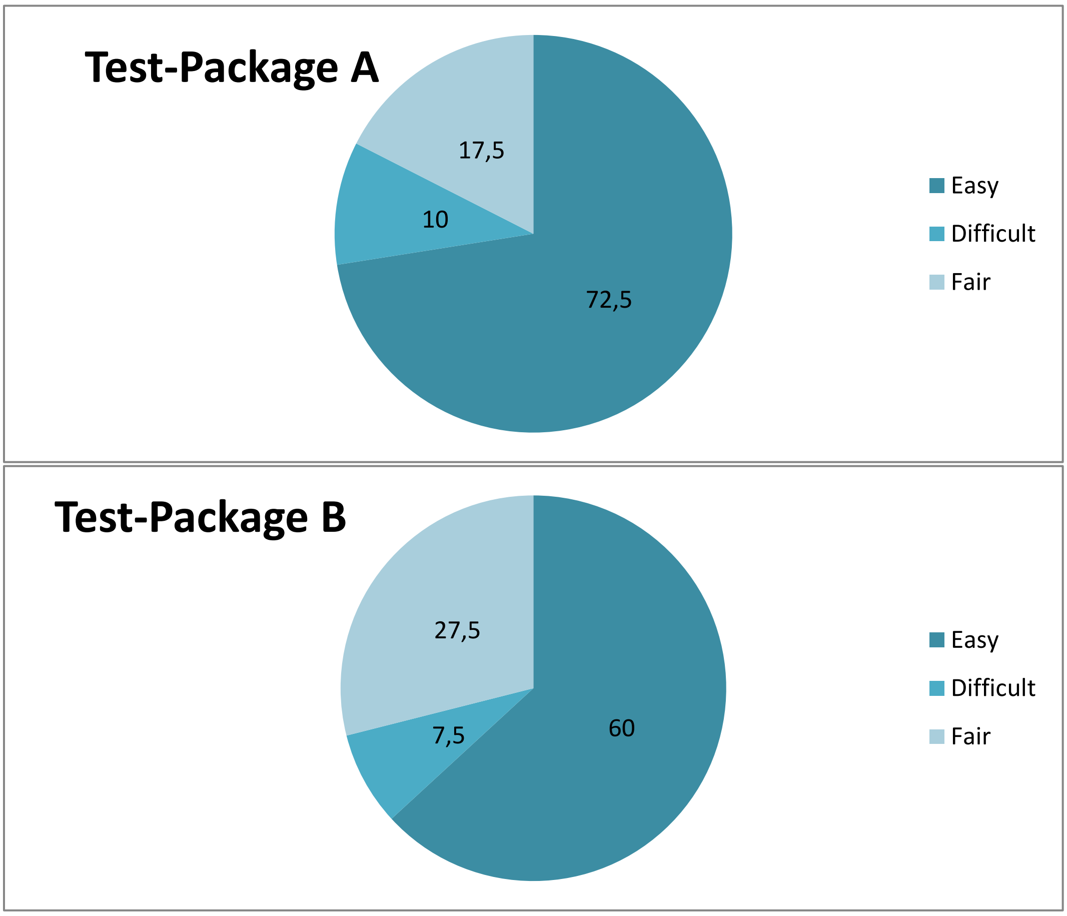
Note:

🟥

The test item number 31 is technically wrong because the instruction is not clear, thus the students cannot answer the question, as the result the teacher gave all correct answer for all options.

Based on the table 4.8 and 4.9, the percentage of the level of difficulty of each test-pack can be shown as the following pie chart:

Figure 4.10.The figure of The Level of Difficulty Percentage (English Final Test of 12th Grade Students of SMAN 1Kedungwaru)



4. Discrimination Power

Discrimination power shows how well a test discriminates between the upper and lower group of the students. The discrimination power of test items can be analyzed by using the following formula:

$$\boxed{DP = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B}$$

(Arikunto, 2012:228)

Where:

DP = Discrimination Power

J =Number of Test-takers

J_A =Total participant of top test-takers

J_B =Total participant of bottom test-takers

B_A = Number of top test takers that have correct answer

B_B =Number of bottom test takers that have correct answer

$P_A = \frac{B_A}{J_A}$ = Proportion of the number of top class answering correctly

$P_B = \frac{B_B}{J_B}$ = Proportion of bottom class answering correctly

The discrimination power can be analyzed by classifying the students into three groups; upper group, middle group, and lower group (for detailed group position, see appendix IV). The researcher took 25% of upper group and 25% of lower group for this analysis and the rest belongs to the middle group which was not used in this analysis.

Arikunto (2012:232) proposed the classification and interpretation of discrimination index of the test items as follows:

Discrimination Index	Classification
0.71-1.00	Excellent
0.41-0.70	Good
0.21-0.40	Satisfactory
≤ 0.20	Poor
Negative value on D	Very Poor

Based on the classification and interpretation of discrimination power proposed by Arikunto, here is the result of discrimination analysis of the test items:

Table 4.11.The Data Presentation of Discrimination Power of Test A

Item	BA	BB	JA	JB	PA	PB	D=PA-PB	Classification
1	0	0	10	10	0	0	0	Poor
2	10	10	10	10	1	1	0	Poor
3	10	9	10	10	1	0.9	0.1	Poor
4	10	10	10	10	1	1	0	Poor
5	10	1	10	10	1	0.1	0.9	Excellent
6	10	10	10	10	1	1	0	Poor
7	10	0	10	10	1	0	1	Excellent
8	10	10	10	10	1	1	0	Poor
9	10	10	10	10	1	1	0	Poor
10	10	9	10	10	1	0.9	0.1	Poor
11	10	9	10	10	1	0.9	0.1	Poor
12	0	0	10	10	0	0	0	Poor
13	10	0	10	10	1	0	1	Excellent
14	10	10	10	10	1	1	0	Poor
15	10	0	10	10	1	0	1	Excellent
16	10	10	10	10	1	1	0	Poor
17	10	1	10	10	1	0.1	0.9	Excellent
18	10	10	10	10	1	1	0	Poor
19	6	10	10	10	0.6	1	-0.4	Very Poor
20	10	6	10	10	1	0.6	0.4	Satisfactory
21	10	10	10	10	1	1	0	Poor
22	10	10	10	10	1	1	0	Poor
23	10	10	10	10	1	1	0	Poor
24	10	10	10	10	1	1	0	Poor
25	10	10	10	10	1	1	0	Poor

26	10	10	10	10	1	1	0	Poor
27	10	10	10	10	1	1	0	Poor
28	10	10	10	10	1	1	0	Poor
29	10	10	10	10	1	1	0	Poor
30	10	10	10	10	1	1	0	Poor
31	10	10	10	10	1	1	0	Poor
32	10	7	10	10	1	0.7	0.3	Satisfactory
33	10	10	10	10	1	1	0	Poor
34	10	10	10	10	1	1	0	Poor
35	10	10	10	10	1	1	0	Poor
36	10	10	10	10	1	1	0	Poor
37	0	1	10	10	0	0.1	-0.1	Very Poor
38	10	1	10	10	1	0.1	0.9	Excellent
39	10	1	10	10	1	0.1	0.9	Excellent
40	10	0	10	10	1	0	1	Excellent

Table 4.12.The Data Presentation of Discrimination Power of Test B

Item	BA	BB	JA	JB	PA	PB	D=PA-PB	Classification
1	10	10	10	10	1	1	0	Poor
2	10	10	10	10	1	1	0	Poor
3	10	10	10	10	1	1	0	Poor
4	10	10	10	10	1	1	0	Poor
5	9	8	10	10	0.9	0.8	0.1	Poor
6	10	6	10	10	1	0.6	0.4	Satisfactory
7	10	10	10	10	1	1	0	Poor
8	9	2	10	10	0.9	0.2	0.7	Excellent
9	5	2	10	10	0.5	0.2	0.3	Satisfactory
10	10	9	10	10	1	0.9	0.1	Poor
11	10	8	10	10	1	0.8	0.2	Poor
12	9	1	10	10	0.9	0.1	0.8	Excellent
13	3	9	10	10	0.3	0.9	-0.6	Very Poor
14	10	10	10	10	1	1	0	Poor
15	10	10	10	10	1	1	0	Poor
16	10	10	10	10	1	1	0	Poor
17	8	8	10	10	0.8	0.8	0	Poor
18	10	10	10	10	1	1	0	Poor
19	10	10	10	10	1	1	0	Poor
20	10	8	10	10	1	0.8	0.2	Poor
21	10	10	10	10	1	1	0	Poor
22	10	10	10	10	1	1	0	Poor
23	8	4	10	10	0.8	0.4	0.4	Satisfactory
24	10	8	10	10	1	0.8	0.2	Poor
25	10	10	10	10	1	1	0	Poor
26	10	7	10	10	1	0.7	0.3	Satisfactory
27	10	8	10	10	1	0.8	0.2	Poor

28	**10**	**8**	**10**	**10**	**1**	**0.8**	**0.2**	**Poor**
29	**7**	**0**	**10**	**10**	**0.7**	**0**	**0.7**	**Excellent**
30	**10**	**7**	**10**	**10**	**1**	**0.7**	**0.3**	**Satisfactory**
31								
32	**10**	**9**	**10**	**10**	**1**	**0.9**	**0.1**	**Poor**
33	**2**	**0**	**10**	**10**	**0.2**	**0**	**0.2**	**Poor**
34	**10**	**5**	**10**	**10**	**1**	**0.5**	**0.5**	**Satisfactory**
35	**10**	**4**	**10**	**10**	**1**	**0.4**	**0.6**	**Satisfactory**
36	**4**	**6**	**10**	**10**	**0.4**	**0.6**	**-0.2**	**Very Poor**
37	**0**	**1**	**10**	**10**	**0**	**0.1**	**-0.1**	**Very Poor**
38	**10**	**1**	**10**	**10**	**1**	**0.1**	**0.9**	**Excellent**
39	**9**	**0**	**10**	**10**	**0.9**	**0**	**0.9**	**Excellent**
40	**10**	**10**	**10**	**10**	**1**	**1**	**0**	**Poor**

Note:

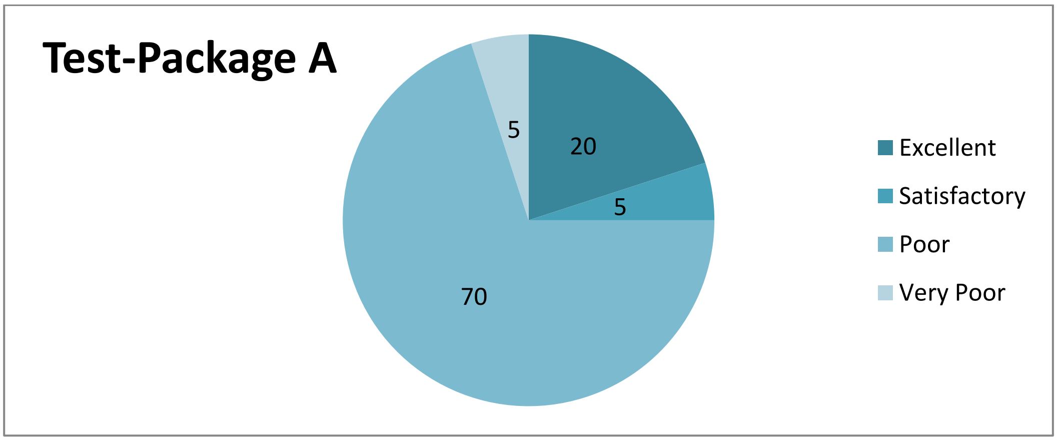
🟥

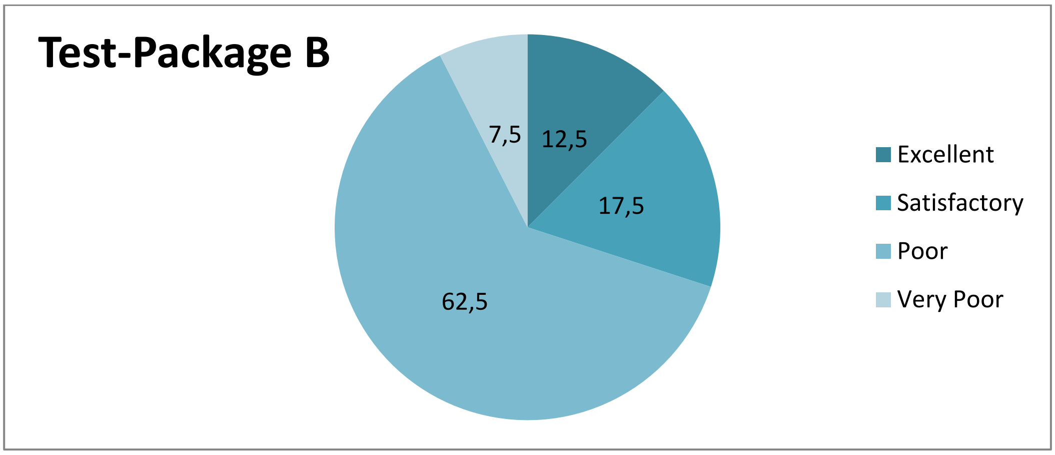
The test item number 31 is technically wrong because the instruction is not clear, thus the students cannot answer the question, as the result the teacher gave all correct answer for all options.

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

Figure 4.13.The Percentage of Discrimination Power

(English Final Test of 12th Grade Students of SMAN 1Kedungwaru)





5. **Distractor Efficiency**

The effectiveness of distractor can be analyzed by finding out the number of students that choose the answers which they believe to be correct, but it was actually wrong answer. A distractor can be said to be well functioned if it has strong power to attract students' believe in choosing the correct answer and if it is chosen by at least 5% of examinees. Here is the table of distractor for each item. The symbol (*) represents the key answer, (+) represents the effective distractor, (-) represents the un-effective distractor, and (O) represents the distractor which must be revised because no one choose it. The effectiveness of distractor of English final test at the 12th grade students of SMAN 1 Kedungwaru is presented in the figure 4.14 below:

Table 4.14.The Effectiveness of Distractor for Each Item(Test PackageA)

Item Number	Options	H (10)	M (20)	L (10)	H+M+L (40)	Percentage	Explanattion
1	A	-	-	-	-	-	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	10	20	10	40	100%	+
2	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	10	20	10	40	100%	*
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
3	A	-	-	-	-	-	O
	B	10	20	9	39	97.5%	*
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	1	1	2.5%	-
4	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	10	20	10	40	100%	*
5	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	10	18	1	29	72.5%	*
	D	-	2	9	11	27.5%	+
	E	-	-	-	-	-	O
6	A	10	20	10	40	100%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
7	A	-	-	-	-	-	O
	B	10	13	-	23	57.5%	*
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	7	10	17	42.5%	+
8	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	2	-	2	5%	-
	D	10	18	10	38	95%	*
	E	-	-	-	-	-	O
9	A	10	20	10	0	100%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O

	E	**-**	**-**	**-**	**-**	**-**	**O**
10	**A**	**-**	**-**	**1**	**1**	**2.5%**	**-**
	B	**10**	**20**	**9**	**39**	**97.5%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
11	**A**	**-**	**1**	**-**	**1**	**2.5%**	**-**
	B	**10**	**19**	**9**	**38**	**95%**	*****
	C	**-**	**-**	**1**	**1**	**2.5%**	**-**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
12	**A**	**-**	**-**	**-**	**-**	**-**	*****
	B	**1**	**1**	**-**	**2**	**5%**	**-**
	C	**9**	**19**	**10**	**38**	**95%**	**+**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
13	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**3**	**10**	**13**	**32.5%**	**+**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**10**	**17**	**-**	**27**	**67.5%**	*****
	E	**-**	**-**	**-**	**-**	**-**	**O**
14	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**10**	**20**	**10**	**40**	**100%**	*****
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
15	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**3**	**1**	**4**	**10%**	**+**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**2**	**9**	**11**	**27.5%**	**+**
	E	**10**	**15**	**-**	**25**	**62.5%**	*****
16	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**1**	**-**	**1**	**2.5%**	**-**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**10**	**19**	**10**	**39**	**97.5%**	*****
	E	**-**	**-**	**-**	**-**	**-**	**O**
17	**A**	**-**	**3**	**-**	**3**	**7.5%**	**-**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**10**	**21**	**1**	**32**	**80%**	*****
	E	**-**	**6**	**9**	**15**	**37.5%**	**+**
18	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**10**	**20**	**10**	**40**	**100%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
19	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**

	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**4**	**9**	**-**	**13**	**32.5%**	**+**
	E	**6**	**11**	**10**	**27**	**67.5%**	*****
20	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**10**	**20**	**6**	**36**	**90%**	*****
	D	**-**	**-**	**4**	**4**	**10%**	**+**
	E	**-**	**-**	**-**	**-**	**-**	**O**
21	**A**	**10**	**20**	**10**	**40**	**100%**	*****
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
22	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**10**	**20**	**10**	**40**	**100%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
23	**A**	**-**	**1**	**-**	**1**	**2.5%**	**-**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**10**	**19**	**10**	**39**	**97.5%**	*****
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
24	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**10**	**20**	**10**	**40**	**100%**	*****
	E	**-**	**-**	**-**	**-**	**-**	**O**
25	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**10**	**20**	**10**	**40**	**100%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
26	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**10**	**20**	**10**	**40**	**100%**	*****
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
27	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**10**	**20**	**10**	**40**	**100%**	*****
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
28	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**10**	**20**	**10**	**40**	**100%**	*****

29	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	10	20	10	40	100%	*
	E	-	-	-	-	-	O
30	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	10	20	10	40	100%	*
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
31	A	-	1	-	1	2.5%	-
	B	10	16	10	36	90%	*
	C	-	3	-	3	7.5%	-
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
32	A	-	-	2	2	5%	-
	B	-	-	-	-	-	O
	C	-	-	1	1	2.5%	-
	D	10	20	7	37	92.5%	*
	E	-	-	-	-	-	O
33	A	10	20	10	40	100%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
34	A	10	20	10	40	100%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
35	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	10	20	10	40	100%	*
36	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	10	20	10	40	100%	*
	E	-	-	-	-	-	O
37	A	-	-	1	1	2.5%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	10	19	9	38	95%	+
	E	-	1	-	1	2.5%	-
38	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	4	9	13	32.5%	+

	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**10**	**16**	**1**	**27**	**67.5%**	*****
39	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**10**	**16**	**1**	**27**	**67.5%**	*****
	E	**-**	**4**	**9**	**13**	**32,5%**	**+**
40	**A**	**-**	**3**	**10**	**13**	**32.5%**	**+**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**10**	**17**	**-**	**27**	**62.5%**	*****

Table 4.12 shows that the effective distractors were shown in option A in number 40; option B in numbers 13 and 15; option C in numbers 12 and 38, option D in numbers 5, 15, 19, 20, 37, and option E in numbers 1, 7, 17, and 39. The un-effective distractors were shown in option A in numbers 10, 11, 23, 31, and 32; option B in numbers 12 and a6; option C in numbers 8, 11, 31, 32; and option E in numbers 3 and 37. While the ommit distractors were shown in option A in numbers 2, 3, 4, 5, 7, 8, 13, 14, 15, 16, 18, 19, 20, 22, 24, 25, 26, 27, 28, 29, 30, 35, 36, 38, and 39; option B in numbers 1, 2, 4, 5, 6, 8, 14, 17, 19, 20, 21, 23, 24, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, and 40; option C in numbers 1, 3, 4, 6, 7, 9, 10, 13, 15, 16, 17, 18, 19, 21, 22, 24, 25, 28, 29, 33, 34, 35, 36, 37, 39, and 40; option D in numbers 1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 14, 18, 21, 22, 23, 25, 26, 27, 28, 30, 31, 33, 34, 35, 38, and 40; and option E in numbers 2, 5, 6, 8, 9, 10, 11, 12, 13, 14, 16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, and 36.

Table 4.15.The Effectiveness of Distractor for Each Item (Test Package B)

Item Number	Options	H (10)	M (20)	L (10)	H+M+L (40)	Percentage	Explanattion
1	A	-	-	-	-	-	O
	B	10	20	10	40	100%	*
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
2	A	-	2	2	4	10%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	10	18	8	36	90%	+
3	A	10	20	10	40	100%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
4	A	10	19	10	39	97.5%	*
	B	-	1	-	1	2.5%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
5	A	-	-	1	1	2.5%	-
	B	1	1	1	3	7.5%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	9	19	8	36	90%	*
6	A	-	-	-	-	-	O
	B	-	4	4	8	20%	+
	C	10	13	6	29	72.5%	*
	D	-	-	-	-	-	O
	E	-	3	-	3	7.5%	-
7	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	10	19	10	39	97.5%	*
	D	-	1	-	1	2.5%	-
	E	-	-	-	-	-	O
8	A	1	2	6	9	22.5%	+
	B	-	-	2	2	5%	-
	C	-	-	-	-	-	O
	D	-	3	-	3	7.5%	-
	E	9	15	2	26	65%	*
9	A	5	8	2	15	37.5%	*
	B	-	-	1	1	2.5%	-
	C	-	1	4	5	12.5%	+
	D	4	11	3	18	45%	+

	E	**-**	**-**	**-**	**-**	**-**	**O**
10	**A**	**10**	**19**	**9**	**38**	**95%**	*****
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**1**	**1**	**2.5%**	**-**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**1**	**-**	**1**	**2.5%**	**-**
11	**A**	**-**	**1**	**1**	**2**	**5%**	**-**
	B	**-**	**-**	**1**	**1**	**2.5%**	**-**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**10**	**19**	**8**	**37**	**92.5%**	*****
12	**A**	**1**	**13**	**9**	**23**	**57.5%**	**+**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**9**	**7**	**1**	**17**	**42.5%**	*****
	E	**-**	**-**	**-**	**-**	**-**	**O**
13	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**7**	**6**	**1**	**14**	**35%**	**+**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**3**	**14**	**9**	**26**	**65%**	*****
14	**A**	**10**	**20**	**10**	**40**	**100%**	*****
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
15	**A**	**10**	**20**	**10**	**40**	**100%**	*****
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
16	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**10**	**20**	**10**	**40**	**100%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
17	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**2**	**-**	**1**	**3**	**7.5%**	**-**
	D	**8**	**20**	**8**	**36**	**90%**	*****
	E	**-**	**-**	**1**	**1**	**2.5%**	**-**
18	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**-**	**-**	**-**	**-**	**-**	**O**
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**10**	**20**	**10**	**40**	**100%**	*****
19	**A**	**10**	**20**	**10**	**40**	**100%**	*****
	B	**-**	**-**	**-**	**-**	**-**	**O**

	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
20	A	-	-	-	-	-	O
	B	10	20	8	38	95%	*
	C	-	-	-	-	-	O
	D	-	-	1	1	2.5%	-
	E	-	-	1	1	2.5%	-
21	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	10	20	10	40	100%	*
	E	-	-	-	-	-	O
22	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	1	-	1	2.5%	-
	D	-	-	-	-	-	O
	E	10	19	10	39	97.5%	*
23	A	1	6	1	8	20%	+
	B	8	13	4	25	62.5%	*
	C	1	1	5	7	17.5%	+
	D	-	-	-	-	-	O
	E	-	-	-	-	-	O
24	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	10	19	8	37	92.5%	*
	D	-	1	2	3	7.5%	-
	E	-	-	-	-	-	O
25	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	10	20	10	40	100%	*
	E	-	-	-	-	-	O
26	A	10	16	7	33	82.5%	*
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	-	4	3	7	17.5%	+
	E	-	-	-	-	-	O
27	A	10	19	8	37	92.5%	*
	B	-	-	1	1	2.5%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	1	1	2	5%	-
28	A	-	1	-	1	2.5%	-
	B	-	-	1	1	2.5%	-
	C	-	-	1	1	2.5%	-
	D	-	-	-	-	-	O
	E	10	19	8	37	92.5	*

29	A	-	2	1	3	7.5%	-
	B	7	5	-	12	30%	*
	C	-	1	2	3	7.5%	-
	D	-	1	1	2	5%	-
	E	2	11	6	19	47.5%	+
30	A	-	-	-	-	-	O
	B	-	-	3	3	7.5%	-
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	10	20	7	27	67.5%	*
31	A	-	5	1	6	15%	*
	B	-	3	4	7	17.5%	*
	C	10	12	4	26	65%	*
	D	-	-	1	1	2.5%	*
	E	-	-	-	-	-	*
32	A	-	-	-	-	-	O
	B	-	-	-	-	-	O
	C	-	-	1	1	2.5%	-
	D	10	20	9	39	97.5%	*
	E	-	-	-	-	-	O
33	A	-	4	9	13	32.5%	+
	B	8	16	1	25	62.5%	+
	C	-	-	-	-	-	O
	D	2	-	-	2	5%	*
	E	-	-	-	-	-	O
34	A	-	1	-	1	2.5%	-
	B	10	19	5	24	60%	*
	C	-	-	-	-	-	O
	D	-	-	-	-	-	O
	E	-	-	5	5	12.5%	+
35	A	-	-	1	1	2.5%	-
	B	10	14	4	28	70%	*
	C	-	5	1	6	15%	+
	D	-	1	4	5	12.5%	+
	E	-	-	-	-	-	O
36	A	-	-	-	-	-	O
	B	6	8	-	14	35%	+
	C	4	11	6	21	52.5%	*
	D	-	-	-	-	-	O
	E	-	1	4	5	12.5%	+
37	A	7	13	6	26	65%	+
	B	-	-	-	-	-	O
	C	-	-	-	-	-	O
	D	3	5	3	11	27.5%	+
	E	-	2	1	3	7.5%	*
38	A	-	3	3	6	15%	+
	B	-	5	6	11	27.5%	+
	C	10	12	1	23	57.5%	*

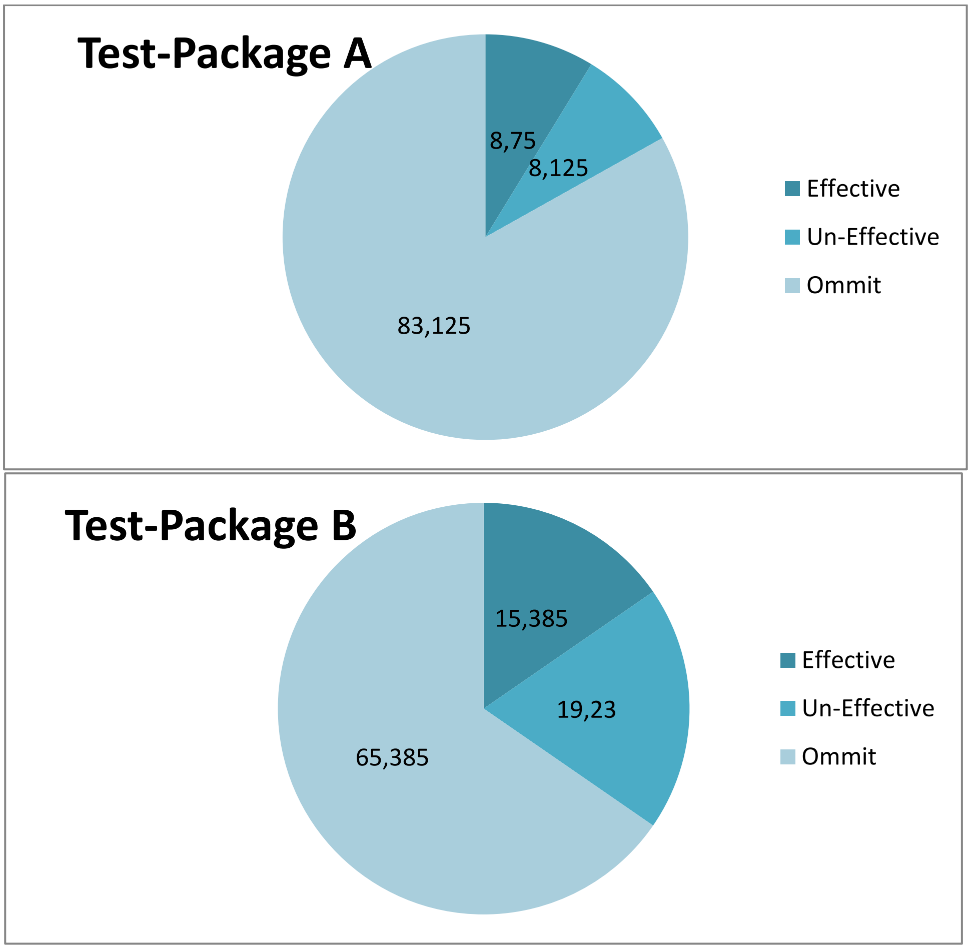
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**
39	**A**	**1**	**1**	**7**	**9**	**22.5%**	**+**
	B	**9**	**15**	**-**	**14**	**35%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**4**	**3**	**7**	**17.5%**	**+**
	E	**-**	**-**	**-**	**-**	**-**	**O**
40	**A**	**-**	**-**	**-**	**-**	**-**	**O**
	B	**10**	**20**	**10**	**40**	**100%**	*****
	C	**-**	**-**	**-**	**-**	**-**	**O**
	D	**-**	**-**	**-**	**-**	**-**	**O**
	E	**-**	**-**	**-**	**-**	**-**	**O**

Figure 4.13 shows that the effective distractors were shown in option A in numbers 8, 12, 23, 33, 37, 38 and 39; option B in numbers 6, 13, 33, 36, and 38; option C in numbers 9, 23, and 35; option D in numbers 9, 26, 35, 37, 39; and option E in numbers 2, 29, 34, and 36. The un-effective distractors were shown in option A in numbers 5, 11, 28, 29, 34, and 35; option B in number 4, 5, 8, 9, 11, 27, 28, and 30; option C in numbers 10, 17, 22, 28, 29, and 32; option D in numbers 7, 8, 20, 24, and 29; and option E in numbers 6, 10, 17, 20, and 27. While the omit distractors were shown in option A in numbers 6, 7, 13, 16, 17, 18, 20, 21, 22, 24, 25, 30, 32, 36, and 40; option B in numbers 7, 10, 12, 14, 15, 17, 18, 19, 21, 22, 24, 25, 26, 32, and 37; option C in numbers 7, 8, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 25, 26, 27, 30, 33, 34, 37, 39, 40; option D in numbers 6, 10, 11, 13, 14, 15, 16, 18, 19, 22, 23, 27, 28, 30, 34, 36, 38, 40; and option E in numbers 9, 12, 14, 15, 16, 19, 21, 23, 24, 25, 26, 32, 33, 35, 38, 39, and 40.

For more detailed information of the effective distractor, here is the data percentage of the effectiveness of distractor.

Figure 4.16. The Percentage of The Effectiveness of Distractor

(English Final Test of 12^{th} Grade Students of SMAN 1Kedungwaru)



2. Discussion

1. Validity

a. Content Validity

Based on the result of the content validity analysis on English final test of the 12^{th} grade students of SMAN 1 Kedungwaru that both the-packages A and B did not cover all of the material stated in the syllabus. It can be seen from table 4.1.which shows that not all materials of each skill in the syllabus were found in the test items. It means that the test items in both test-packages did not represent the overall material taught by the teacher. 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 supposes to measure”. Thus it is very important for the test designer to consider the content validity of the test items because the result of the test items, later, will be used as the representative of the students’ achievement and if the test items do not have good content validity, it is impossible to make a use of the test result.

From the table 4.2, it can be also shown that the proportion of the content validity represented in the test items was not fair in which most of the test items were testing reading skill. The percentage shows that 77.5% is testing reading, 17.5% is testing writing, 5% is testing speaking and 0 % for listening skill. This percentage was obtained for both test-packages A and B. It leads to be lack of content validity because the test items only tested three of four skills of language should be tested.

In the problem on the content validity of a test, the test-designer should make sure that all of the material stated in the syllabus has been covered in the test items. The test designer can modify the form of the test in order to cover all of the material taught to the students. In order to test listening skill, for example, the test-designer can give listening test which asks the students to listen on a certain recording then asked them to answer the questions in the form of multiple-choice test, short answer questions, or may be true false questions. This test can be done at the same time or different time of doing the written test.

While for writing and speaking test, the content validity seemed to be not really fair since the material of both skills were not all covered in the test items, so that the test designer should give more proportion of both skills in the test items, thus the proportion of the content validity of the test items will be balance, or it will be better if the test designer also conducts a special test for both skills because these skills needs practice to know how far the students master the material of both skill.

In order to have good content validity, the test maker needs a specification of the skills or subjects that is meant to cover in the test and the test makers must ensure that the specification they made is based on the principled selection of elements for inclusion in the test (Huges, 1989:22). In addition, test maker should also haveattempt to 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 components skill and area which the test maker wishes to include in the assessment".

b. Construct Validity

The technique used in English final test of the 12th grade students of SMAN 1 Kedungwaru was multiple-choice test which assessed only three of four skills should be assessed including speaking, reading, and writing. Multiple-choice test is an appropriate form of test especially for reading skills. Multiple-choice test is considered to be the best form of testing reading notably for reading passage. Madsen (1983:83) states that one of the best methods in testing reading passage is multiple-choice test. Multiple-choice test is sufficient since more than one passage will appear on a single test.

In English final test of the 12th grade students of SMAN 1 Kedungwaru, the form of reading test was already tested both micro and macro skills of reading skill. The test items had already tested the micro skills underlying reading skill like identifying referents of pronouns, using context to guess meaning of unfamiliar words, understanding relations between parts of the test. In addition, the test had also tested the macro skill of reading like scanning text to locate specific information, skimming text to obtain general idea, identifying stages of argument, and identifying examples presented in support of an argument. The test form used in English final test of the 12th grade students of SMAN 1 Kedungwaru was appropriate enough for the students' level. Thus, the test items testing reading were acceptable for both test-packages A and B.

Previously, it was explained that the multiple-choice test was appropriate for reading test, however, this kind of testing technique was not appropriate enough for testing writing moreover for testing speaking. Because both skills are productive skills and they require students to practice their ability in producing or expressing ideas through writing and speaking.

The first is about speaking test, the students only asked to choose the response for a certain dialogue and categorize the dialogue whether the dialogue belongs to expressing of forgiving or promising and etc. Whereas, speaking proficiency actually not only deals with ability of responding to a certain dialogue but with the ability of pronouncing words or even sentence, mastering grammar, vocabulary, fluency and appropriateness of expression are equally important to be evaluated by the teacher. It is impossible to know the ability of the students in pronunciation, fluency and also appropriateness of expression by having multiple-choice test. Thus it is necessary for the teacher to conduct a speaking test which asks students to have speaking practice.

The second one is writing test. According to Madsen (1983: 101) that there many aspects can be evaluated in writing test: mechanics (including spelling and punctuation), vocabulary, grammar, appropriate content, diction, and rhetorical matters of various kinds (organization, cohesion, and unity). Those aspects of writing skills cannot be evaluated only by having multiple-choice test. Thus, teacher also needs to conduct a writing test in order to make the students practice their ability in writing.

Similar with the reading skill, multiple-choice test is actually appropriate enough for testing listening skill, however, the problem of the construct validity of the English final test of the 12th grade students of SMAN 1 Kedungwaru in testing listening is that the test items did not test the underlying skill of listening because the teacher did not ask the students to listen on a certain recordings. Hughes (1989:134-135) stated that testing listening must involve testing macro and micro skill of listening. The macro skills of listening include; listening for specific information, obtained gist of what is being said or listened, and following instruction; and the micro skill of listening include the ability of the students in interpreting the intonation pattern and recognition of structure function. That's way, the test items of SMAN 1 Kedungwaru for both test-packages A and B were lack of construct validity because most of the test items tested the underlying ability of reading skill.

To make good construct validity, it is supposed to use appropriate or even various technique of testing to assess the skills of language. The teacher should not use the single form of test, multiple-choice test, to test all of language skills and components. In addition, Heaton (1983: 161) explained that "……. 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".

A test is said to have construct validity if it only measures 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".

2. Reliability

The result of reliability coefficient of English final test of the 12th grade students of SMAN 1 Kedungwaru for test-package A was 0.72 and 0.48 for test-package B. The reliability coefficient of test-package A is considered to be fair, while the reliability coefficient of test-package B is considered to be low because it is less than 0.5. Reliability is one of the five principles of language testing proposed by Brown. Thus, it is very necessary for the test designer to know the reliability of the test items they made. A good test can be considered to be a valid test, if it is also reliable because a reliable test is a test that can produce correct or true score which can be trusted. Reliability is thus measure of accuracy, consistency, dependability or fairness of scores resulting from administration of particular examination.

In this study, the researcher found that there is a significant difference found in the reliability of test-packages A and B with the reliability coefficient of test-package A was higher than test-package B those are 0.72 for test-A and 0.48 for test-B, whereas both test-packages were administered in the same class and in the same level. This condition should not be happened because both test-packages were used interchangeably to the students. So, the test-maker must ensure that the reliability coefficient of both test-packages is equal because it was not fair for the students if the reliability coefficient of both test-packages was not equal.

The difference of the reliability coefficient between these test-packages may be caused by some factors like the condition of the test-takers, classroom situation or any other factors affecting the concentration of the test-takers in doing the test.

Brown (1996: 188-189) proposes errors of measurement; some issues that may affect the reliability coefficient of a test. First, the issues due to the environment: location, ventilation, space, noise, lighting and weather. Second, the issues due to administration procedures: direction, equipment, timing and mechanics of testing. Third, the issues due to the test-takers: health, fatigue, physical characteristics, motivation, emotion, memory, concentration, forgetfulness, impulsiveness, careless, comprehension of direction, guessing, and chance knowledge of item content. The next is the issues due to scoring procedure: errors in scoring, subjectivity, evaluator biases, and evaluator idiosyncrasies. The last is the issues due to the test and test items: test booklet clarity, answer sheet format, particular sample of items, number of items, item quality and test security.

In order to avoid the measurement errors, there are some alternatives to create more reliable test. Hughes (1989:36-43) suggests ways of achieving more reliable 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 technique
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

Creating different test items for the same group of students is not easy, thus the test-designer should consider those alternatives way in creating more reliable test in order to create the different test-packages to be administered for the same group which has equal reliability coefficient for each test.

3. Level of Difficulty

The percentage of the level of difficulty of English final test of the 12th grade students of SMAN 1 Kedungwaru for both tests-packages A and

B shown in figure 4.10 showing that the difficulty level percentage of test package A: 72.5% were easy items, 17.5% were fair items, and 10% were difficult items. While the difficulty level percentage of test-package B: 60% were easy items, 27.5% were fair items and 7.5% were difficult items. From the percentages above, it can be seen that the proportion of difficulty level of both test-packages was not equal in which test-package A had more difficult test items rather than test-package B. It should not be happened because both-testpackages were used interchangeably to the same group of the students; therefore the teacher or test-maker must ensure that the proportion of the difficult, fair and also easy test items for both test-packages must be equal, so that both-test-packages will be fair.

The items of both test-packages must be in appropriate level of difficulty for the students to whom the test is administered. The test designer must make a test which has indices of difficulty level no less than 0.31 and no greater than 0.70 and if the test-designer is intended to create two different test-packs to be administered for one class at the same level, thus the test-designer must ensure that the difficulty level of both test-packs must be at the same level, or if it is impossible, at least the difference is not too far. Thus, it is desirable for the test-designer to have most items in the 0.31-0.70 range of difficulty. Too difficult or too easy items are not effective to use for discriminating the students.

The difficulty level of the test items has the relationship with the arragement of the test items and the arragement of the test items itself gives certain effect on the students' confidence in doing the test. Djiwandono (2008:220) states that giving difficult question which makes the students think harder and consume more time to answer at the beginning numbers will lead to give bad effect for the students because, they will feel inferior and afraid in doing the difficult items in the test and it also affects to the next questions, so the students will also be affraid and unconfidence in doing the test even though the test items is actually not as difficult as the previous questions.

The difficult test items must be arraged in the last numbers in order to make the students fell confident in doing the test because they have done the beginning numbers of the test well and easily. So, if the students find trouble in doing the test items in the last numbers, it will give no effect to the students because they have done the previous numbers well. In conclusion, the test-designer must also consider about the arrangement of the test items which should be arranged by the easy items at the beginning numbers, the fair items at the middle numbers, and the difficult items at the last numbers of the test items.

4. Discrimination Power

The result of discrimination power analysis was shown in figure 4.11 showing that most of the test items cannot give the information about the difference of the students' ability in answering the test because mostly the test items have poor discrimination power. For test-package A,the test items having poor discrimination power were shown in numbers 1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 14, 16, 18, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, and 36. While for test-package B, the test items having poor discrimination power were shown in numbers 1, 2, 3, 4, 5, 7, 10, 11, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 27, 28, 32, 33, and 40. Those test items were categorized into poor discriminator because the interval of the upper students and the lower students answered the questions correctly is little, the interval is around 0-3. Thus these test items are still acceptable but need to be improved in order to achieve the criteria of good or satisfactory distractor.

Next, the test items having satisfactory or functioned discriminatorfor test-package A were only shown in the test items number 20 and 32, while for test-package B were shown in the test items numbers 6, 9, 23, 26, 30, 34, and 35. In addition, the test items having excellent discriminator in test-package A were shown in the test items numbers 5, 7, 13, 15, 17, 38, 39, and 40, while in test-package B were shown in numbers 8, 12, 29, 38, and 39. These items were categorized into satisfactory and excellent distracriminator because they had the information about the differences in the students' performance especially for the upper and lower group and they can also discriminate the students' ability, therefore, the test-designer can keep saving those items in order to be administered in the next test.

Besides the poor, satisfactory and excellent distracriminators, the researcher also found the negative result of the discrimination analysis of

English final test of the 12th grade students of SMAN 1 Kedungwaru which were shown in the test items numbers 19 and 37 in test-package A, and numbers 13, 36, and 37 in test-package B. The negative result of the discrimination analysis shows that the test items have very poor discrimination power because the students from upper group who are supposed to answer the question correctly answered the questions incorrectly, on the contrary, the students from lower group who are supposed to answer the questions incorrectly answered the questions correctly, or it can be said that the numbers of the upper group answered the questions correctly was fewer than the lower group. Thus, these kinds of items must be all removed.

The percentages of the discrimination power analysis of both test-packages showed different percentages of 70% of poor discrimination power for test-package A and 62.5% of poor discrimination power for test-package B (the detailed difference can be seen in the figure 4.13). This difference should not be found when two kinds of test-packages were administered interchangeably to the same group. Thus, it is a must for the test-maker to create two different test-packages with the same proportion of the discrimination power if the two test-packages are used interchangeably to the same group of the students.

Discrimination is one of the important features of good test. It is the ability of an item to discriminate among the difference candidates, reflect the difference performance of the individuals in a certain group and distinguish among the students who have high ability in responding the questions correctly and those who have lower ability in responding the questions correctly. The higher the discrimination index of the test items is, the better it is.

Sudjiono (1996: 408) states that following up after analyzing the discrimination power of a certain test must be done by the teacher or test-maker in order to revise the test items. The follow up proposed by Sudjiono are as follows:

a. The items which have good discrimination power; satisfactory and excellent classification; should be kept in item test bank, so that it can be used later.

b. The items which are categorized into the poor distractor should be revised and then used later.

c. The very poor discriminator of the test items then must be dropped or removed because it cannot be used later.

5. The Effectiveness of Distractor

A typical multiple-choice test consists of a question, referred to as the stem, and a set of two or more options that consist of possible answers; one correct answer and distractors; to the question. All of the distractors or incorrect options should actually attract the students' attention in choosing the correct answer. Preferably, each distractor should be chosen by a greater proportion of the lower group than that of the upper group. The effectiveness of distractor has inseparable relationship with the discrimination power of the test items. If the distractors of the test items are not effective, definitely the test items will so have low discrimination power because the lower group of the students will be able to answer the questions correctly and easily.

Figure 4.12 and 4.13 show the result of the analysis on the effectiveness of distractor of the English final test of the 12th grade students of SMAN 1 Kedungwaru for both test-packages A and B. The result shows that the effective distractors were shown in option A in number 40; option B in numbers 13 and 15; option C in numbers 12 and 38, option D in numbers 5, 15, 19, 20, 37, and option E in numbers 1, 7, 17, and 39 for test-package A; while in test-package B, the effective distractors were shown in option A numbers 8, 12, 23, 33, 37, 38 and 39; option B in numbers 6, 13, 33, 36, and 38; option C in numbers 9, 23, and 35; option D in numbers 9, 26, 35, 37, 39; and option E in numbers 2, 29, 34, and 36. They are categorized into effective distractor because there at least 5% of the students chosen those distractors, so that the effective distractors should be kept and they are still able to be used for the next test.

Besides that, the researcher also found that there are un-effective distractors which were shown in option A in numbers 10, 11, 23, 31, and 32; option B in numbers 12 and a6; option C in numbers 8, 11, 31, 32; and option E in numbers 3 and 37 for test-package A; and for test-package B they are found in option A in numbers 5, 11, 28, 29, 34, and 35; option B in number 4, 5, 8, 9, 11, 27, 28, and 30; option C in numbers 10, 17, 22, 28, 29, and 32; option D in numbers 7, 8, 20, 24, and 29; and option E in numbers 6, 10, 17, 20, and 27. The un-effective distractors are those which are chosen by less than 5% of the students or examinees. Thus, the un-effective distractor should be revised in order to reach the criteria of good distractor because the quality of the distractor will affect the discrimination power of the test item.

The other distractors from both test-packages A and B which were not mentioned above are categorized as the omit distractors because those distractors did not attract students' attention in choosing the correct answer or nobody chose those distractors. Therefore, this kind of distractor must be removed or deleted.

The percentage of the distractor efficiency analysis for both test-packages also showd different result where test-package A had the higher proportion of the omit distractor than test-package B with the percentage of 83.125% of test-package A and 65.385% for test-package B. However, this difference should not be found when two kinds of test-packages were administered interchangeably to the same group. Thus, it is a must for the test-maker to create two different test-packages with the same proportion of the distractors if the two test-packages are used interchangeably to the same group of the students.

The distribution of distractors means the distribution of alternative answers. The importance of calculating it is to know the students' answers. Arikunto (2012: 238) points out that a distractor can be said to have functioned well when it is chosen by at least 5% of the total examinees. If the index of this is 0, thus the distractor should be discarded or eliminated with a more effective option. In addition, some distractors may be too appealing and causing the items to be too difficult. Very often items which have been rejected as having inappropriate difficulty, discrimination power, or omit distractor can be redeemed by the revision of one or two response options.

In order to write good distractors or possible answers in the multiple choice test, Haladyna (2004: 99) suggests how to write good options; either distractors or answer key as follows:

1. Develop as many effective options as the test maker can
2. Vary the location of the correct answer according to the number options. Assign the position of the correct 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 distractors plausible.
7. Avoid negative words such as not or except.
8. Avoid options that give clues to the right answer.
9. Use typical errors of students when writing distractors.

CHAPTER V

CONCLUSION AND SUGGESTION

A. Conclusion

After analyzing the obtained data about students' answer sheets and the test items on the English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215, the researcher deduced five conclusions related to the test item analysis as follows:

1. English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was lack of content validity because the test items did not represent all of the materials stated in the syllabus, in addition the test items also did not test four skills provided in syllabus completely. The percentage of the skills being tested was 0% items for testing listening in both test-packages A and B, 5% for testing speaking in both test-packages A and B, 77.5% for testing reading in test-packages A and 80% for test-packages B. 17.5% for testing writing in test-package A and 20% for test-package B.

   In addition, the English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was also lack of construct validity to test some skill of four skills to be tested. The test item was lack of construct validity to test listening, speaking and also writing; but the test was good in the construct validity to test reading skill because multiple-choice test was appropriate to test reading skill, however, multiple-choice test was not appropriate for testing speaking and writing because these skills need to be practiced in order to know the students' proficiency and evaluate any aspects related to these skills. While, the multiple-choice is actually appropriate for listening test, but it should be supported with the recording in order to test the students' ability in listening to a certain sounds.

2. The reliability coefficient of test-packages A and B shows different result in which the reliability coefficient of test-packages A is higher than test-package B. the reliability coefficient of test-package A was 0.72. It means that the reliability of the test was categorized as fair reliability test. So that the test items are acceptable to use. However, the coefficient for test-package B is lower, 0.48. It means that the reliability coefficient of test-package B was categorized as low reliability test and it is not acceptable to use for testing the students.
3. The percentage of the level difficulty of the English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was 72.5% of easy test items for test-package A and 60% for test-package B; 17.5% of fair items for test-package A, and 27.5% for test-package B; and 10% of difficult items for test-package A and 7.5% for test-package B. As it was shown that both test-packagess were dominated by the easy items and too easy items were not good for the students and it also related to the discrimination power of the test items.

The lower difficulty level of the test items is, the lower discrimination power of the test items is.

4. The discrimination power of the test items of English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was low because mostly the test items of both test-packages A and B were dominated by the items which have poor discrimination power represented in the percentage of 70% of poor discrimination for test-package A and 62.5% for test-package B.
5. The percentage of the distractor analysis on the English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was mostly dominated by the omit distractors for both test-packages A and B and the effectiveness of distractors has positive correlation with the discrimination power of the test, thus if the test has bad distractors, thus the discrimination power of the test must be low.

On the basis of the conclusion above, it can be drawn a general conclusion that the quality of the English final test of the 12th grade students of SMAN 1 Kedungwaru at the first semester in academic year 2014/215 was not good in term of its validity; content and construct validity for both test-packages A and B, the reliability coefficient for test-package B, the difficulty level for both test-packages, the discrimination power for both test-packages, and the distractor efficiency for both test-packages. Those aspects of the test must be revised for the improvement.

B. Suggestion

Based on the research findings described in the previous chapter, some suggestions were given to the english teacher, test-maker, and other researcher.

1. The teacher

    It is suggested for the English teacher in doing evaluation that the teacher should not rely only on the result of the final test to know the students language mastery level because commonly the final test is in the form of multiple-choice test and this form is not appropriate for testing some skills of English. Thus, the teachers should also assess the students' progress doing teaching and learning process by using authentic assessment.

    In addition, in creating a test, the teachers or test-makers must make sure the quality of the test-pack they made in term of its validity, reliability, difficulty level,discrimination power, and distractor efficiency in order to create good test instrument. The teachers or test-makers should try out the test before administering it to the students. It is beneficial for the teachers or test-makers in order to know the weaknesses of the test they made. Thus, doing evaluation on the quality of the test items is also necessary especially for the teachers in order to know the quality of the test itself.

2. Other researchers

   It is expected for further researchers that if they want to continue this research they should not just analyze and describe the quality of the English final test, but they should also interview the test-maker the way they create the test or the students, if it is necessary, in order to get deep information on it.

REFERENCES

Allison, D. 1999. *Language Testing: An Introductory Course.* Singapore: Singapore University Press

Arikunto, S. 2012. *Dasar-dasarEvaluasiPendidikan.*Jakarta: BumiAksaraPress

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

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing.* London: Oxford University Press.

Brown, H.D. 2000. *Teaching by Principle: An Interactive Approach to Language in Pedagogy.* White Plains, NY: Pearson Education

Brown, H. D. 2004. *Language Assessment: Principles and Classroom Practices.* White Plains, NY: Pearson Education

Bumagat. W.C. 2004. *The Function of Measurement and Evaluation in Improving Instruction.* Turod Elementary School Solana North Distric. Solana, Cagayan. http://www.bilaterals.org/IMG/doc/the function of measurement and evaluation in improving instruction.doc

Cohen, et al. 2007. *Research Method in Education*. New York: Routledge

Djiwandono,M.S. 2008. *TesBahasa*. Jakarta: Indeks

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

Fulcher, Glenn. 2010. *Practical Language Testing.* London: Hodder Education, AN Hachette UK Company

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

Henning, Grant. 1987. *A Guide to Language Test*. London: Longman

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge : Cambridge University

Madsen. H. S. 1983. *Technique in Testing.* New York: Oxford University Press

Osterlind, J.S. 2002. Constructing Test Items; Multiple Choice, Constructed-Response, Performance, and Other Forms. New York: Kluwer Academic Publishers

Salwa, A. 2012. *The Validity, Reliability, Level of Difficulty and Appropriatness to The Curriculum of English Test*. Semarang: Universitas Diponegoro Semarang

Sudjiono, Anas. 1996. *PengantarEvaluasiPendidikan.* Jakarta: Raja GrafindoPersada

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

Tuckman, B. W. 1975. *Measuring Educational Outcomes Fundamentals of Testing.* New York: Harcourt Brace Javanovich Inc.

Valette, R.M. 1967. *Modern Language Testing.* New York: Harcourt Brace Jovanovich Publishers, Inc.

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