AN ANALYSIS OF ENGLISH SEMESTER TEST ITEMS BASED ON THE CRITERIA OF A GOOD TEST FOR THE FIRST SEMESTER OF THE FIRST YEAR OF SMK NEGERI 1 GEDONG TATAAN IN 2012/2013 ACADEMIC YEAR

(A Script)

By

NUR SARTIKA PUTRI

TEACHER TRAINING AND EDUCATION FACULTY
LAMPUNG UNIVERSITY
BANDAR LAMPUNG
2015

ABSTRACT

AN ANALYSIS OF ENGLISH SEMESTER TEST ITEMS BASED ON THE CRITERIA OF A GOOD TEST FOR THE FIRST SEMESTER OF THE FIRST YEAR OF SMN NEGERI 1 GEDONG TATAAN IN 2012/2013 ACADEMIC YEAR

By

NUR SARTIKA PUTRI

The problem of the reseach was focused on the quality of test items used in semester exams. The objectives of the reseach were intended to determine the quality of English semester test items whether or not fulfilled the following criteria of a good test: (1) face validity, (2) content validity, (3) construct validity, (4) reliability, (5) discrimination power, (6) level of difficulty, and (7) the quality of options.

The method used in the research was descriptive analytic. The data were collected through test and questionnaire. The test paper obtained, together with the questionnaire, was analyzed to investigate face validity and construct validity of the test. The test paper, together with the materials in English curriculum was analyzed to investigate content validity. The students answer sheets were analyzed to investigate reliability, discrimination power, level of difficulty, and the quality of options.

Based on the findings, the results of the analysis proved that the English semester test items had (1) good face validity since the students answered 69,3 % yes-answers of the questionnaire given, (2) good content validity since fifty objective items test represent the subject matter content available in English Curriculum 2006, (3) good construct validity since respondents gave 75% yes-answer correctly, (4) has low reliability (r = 0.07), it implied that the items test needed some revisions, (5) the result of discrimination power, there were 13 poor items included in 3 items were negative discrimination, 24 satisfactory items, 12 good items, and there was only one excellent item on this achievement test. (6) The level of difficulty of the test consisted of: 17 easy items, 16 difficult items, and 17 average items. If it is seen from both DP and LD, there were 3 items that be discarded, 25 items should be revised and 22 items were acceptable. (7) The test has the quality of options by using ITEMAN: 29 alternatives have rejected, 60 alternatives accepted, and 111 alternatives revised from 200 distracters.

Key words : *test items analysis, criteria of a good test, iteman*

AN ANALYSIS OF ENGLISH SEMESTER TEST ITEMS BASED ON THE CRITERIA OF A GOOD TEST FOR THE FIRST SEMESTER OF THE FIRST YEAR OF SMK NEGERI 1 GEDONG TATAAN IN 2012/2013 ACADEMIC YEAR

NUR SARTIKA PUTRI

A Script
Submitted in a partial Fulfillment of
The Requirement for S-1 Degree
in
The Language and Arts Department of
Teacher Training and Education Faculty

TEACHER TRAINING AND EDUCATION FACULTY
LAMPUNG UNIVERSITY
BANDAR LAMPUNG
2015

Research Title : **AN ANALYSIS OF ENGLISH SEMESTER TEST ITEMS BASED ON THE CRITERIA OF A GOOD TEST FOR THE FIRST SEMESTER OF THE FIRST YEAR OF SMK NEGERI 1 GEDONG TATAAN IN 2012/2013 ACADEMIC YEAR**

Student's Name : Nur Sartika putri

Student's Number : **0743042028**

Department : **Language and Arts Education**

Study Program : **English Education**

Faculty : **Teacher Training and Education**

APPROVED BY

Advisory Committee

Advisor	Co-Advisor
H. M. Ujang Suparman, M.A., Ph.D.	**Drs. Ramlan Ginting Suka, M.Pd.**
NIP 19570608 198603 1 001	NIP 19570721 198603 1 003

The Chairperson of
The Department of Language and Arts Education

Dr. Mulyanto Widodo, M.Pd.
NIP 19620203 198811 1 001

ADMITTED BY

1. Examination Committee

   Chairperson : **H. M. Ujang Suparman, M.A., Ph.D.** ______________

   Examiner : **Prof. Dr. Cucu Sutarsyah, M.A.** ______________

   Secretary : **Drs. Ramlan Ginting Suka, M.Pd.** ______________

2. The Dean of Teacher Training and Education Faculty

Prof. Dr. Hi. Bujang Rahman, M.Si.
NIP 19600315 198503 1 003

Graduated on : **November 02nd, 2015**

SURAT PERNYATAAN

Sebagai civitas akademik Universitas Lampung, saya yang bertanda tangan dibawah ini,

Nama	:	Nur Sartika Putri
NPM	:	0743042028
Judul Skripsi	:	An Analysis of English Semester Test Items Based on the Criteria of a Good Test for the First Semester of the First Year of SMK Negeri 1 Gedong Tataan in 2012/2013 Academic Year
Program Studi	:	Pendidikan Bahasa Inggris
Fakultas	:	Keguruan dan Ilmu Pendidikan

Dengan ini menyatakan bahwa

1. Karya tulis ini bukan saduran/terjemahan, murni gagasan, rumusan, dan pelaksanaan penelitian/implementasi saya sendiri tanpa bantuan dari pihak manapun kecuali arahan Pembimbing Akademik dan Narasumber di organisasi tempat melaksanakan riset.
2. Dalam karya tulis ini terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
3. Pernyataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai dengan norma yang berlaku di Universitas Lampung.

Bandar Lampung, November 2015
Yang membuat pernyataan,

Nur Sartika Putri
0743042028

CURRICULUM VITAE

The writer’s name is Nur Sartika Putri. She was born in Kotabumi, Lampung Utara, April 11, 1988 as out of nine daughter of *Bapak* Hi. Syafli Nafis and *Mama* Hj. Bainah Heriyati.

She went to the formal institution for the first time in 1992 at TK Tunas Harapan Depak, Kotabumi and graduated in 1994. After that, she was registered at Elementary school, SD Negeri 10 Candimas, and graduated in 2000. Then, she continued her study at Junior High School, SLTP Negeri 7 Kotabumi, and graduated in 2003. After graduating from Junior High School, she was enrolled at SMA Negeri 1 Kotabumi, and graduated in 2006. In the following year, she was registered as a newly accepted student at the English Study Program, Language and Art Education Department of Lampung University.

She joined the Teaching Practise Program as one of the requirements for FKIP students at SMK Negeri 4 Bandar Lampung from July 20 until September 29, 2011.

DEDICATION

All praises and gratitude are only to Alloh SWT, for all His tremendous blessings.

I’d proudly dedicate this script to :

my beloved father and mother, thanks for their endless love, support, and prayer.

my beloved sisters and brothers, thanks also for their guidance, spirit, and support, I love them all with my heart.

my lovely nephews and nieces, for brightening my days.

my beloved husband to be, who will be my guardian heart whom I will share my life with.

my close-friends, for giving me the courage to finish this script.

my friends in SMK Negeri 1 Gedong Tataan.

my friends in English Department Unila.

my dear lectures, for their greatest knowledge.

my beloved almamater, University of Lampung.

MOTTO

لا حول و لا قوة الا بالله
lâ hawla wa lâ quwwata illâ bi Allâh
“There is no power or strength except through Alloh”

“Think carefully. Choose wisely.”
(Nur Sartika Putri)

ACKNOWLEDGEMENT

Alhamdulillahirobbil'alamin, Praise to the Most Gracious and Merciful, Alloh SWT for His tremendous blessing that the writer is able to finish her script with the title *"An Analysis of English Semester Test Items Based on the Criteria of a Good Test for the First Semester of the First Year of SMK Negeri 1 Gedong Tataan in 2012/2013 Academic Year"*. This script is submitted as a compulsory fulfillment of the requirements for S1 degree of English Study Program, The Language and Arts Department, The Faculty of Teacher Training and Education, Lampung University.

It is important to know that this script would have never come into existence without any support, encouragement, and assistance by several dedicated persons. Here are the writer would like to address her gratitude and respect to:

1. H. Ujang Suparman, S.Pd., M.A., Ph.D., as the writer's supervisor, for his willingness to give assistance, ideas, encouragement, and scientific knowledge within his time during the script writing process.
2. Drs. Ramlan Ginting Suka, M.Pd as the writer's co-supervisor, for his kindness, suggestion, and patience in guiding the writer finishing this script.
3. Prof. Dr. Cucu Sutarsyah, M.A., as the writer's examiner for his valuable suggestions and ideas to make this script more valuable.
4. Dra. Hartati Hasan, M.Hum., as the writer's academic advisor.
5. All lectures of English Program, who have contributed their guidance and knowledge for the writer.
6. Dr. Ari Nurweni, M.A. as the Head of English Study Program.
7. Dr. Mulyanto Widodo, M.Pd. the Chairperson of Language and Arts Education Department.
8. Dr. H. Bujang Rahman, M.Si, the Dean of Teacher Training and Education Faculty.
9. Drs. Sutomo,M.M, as the headmaster of SMK Negeri 1 Gedong Tataan for allowing her to undertake the research, and to Mrs. Susilawati, S. Pd., as the English teacher of the school for giving the opportunity and time to conduct this research. Her thanks also go to the students of Computer Department and Automotive Department who welcomed her into their environment and so willingly became involved in the research.

10. My beloved parents, *Bapak* Hi. Syafli Nafis and *Mama* Hj. Bainah Heriyati, thanks for their endless love, support, patience, and pray, may the only Alloh SWT pay their love and devotion back within His never ending blessing. Thanks for everything in my life. I love you so much.
11. My beloved sisters and brothers: *Abang* Eka, *Ucik* Heri, *Ayuk* Mega, *Ayuk* Fatma, *Ayuk* Susi, *Ayuk* Tina, *Ayuk* Linda, *Abang* Budi, *Abang* Wan, who always give me guidance, support, and encourage to face everything.
12. Special thanks to my mood booster, Andan Susanto S.Kom, who has given his support, love, guidance, suggest, pray, and patience. Thanks for being such a great companion along the way in finishing this scriptsweet and always beside me even in down and up.
13. My close friends: Silvia Agustin S.Pd, Vera Welianti S.Pd, Mesi Astriani S.Pd, Lina Fitri Yani A.Md, Ruanitatur Irmareti S.Pd, Annisa Kencana W S.Pd, Dwi Paramitha S.Pd, Desi Putriani S.Pd, Yuyun Fadillah S.Pd, and Elvina Natalia S.Pd, etc. Thanks for all the happy times, support, and unforgettable best helps during my studying period and especially when I did my research and wrote this script.
14. My friends in English Study Program and the other Study Program, who can’t be mentioned directly who has contributed in completing this script.
15. My friends in SMK Negeri 1 Gedong Tataan.

Last but no least, the writer truly realizes that this research report might not be perfect yet; thus, constructive input is accordingly expected to make the paper in the future.

Bandar Lampung, November 2015
The writer,

Nur Sartika Putri

TABLE OF CONTENTS

Page

ABSTRACT .... **i**
APPROVAL .... **ii**
ADMISSION .... **iii**
SURAT PERNYATAAN .... **iv**
CURRICULUM VITAE .... **v**
DEDICATION .... **vi**
MOTTO .... **vii**
ACKNOWLEDGEMENT .... **viii**
TABLE OF CONTENT .... **ix**
LIST OF TABLE .... **xii**
LIST OF APPENDICES .... **xiii**

I. INTRODUCTION

1.1 Background of the Problems 1
1.2 Identification of the Problems 5
1.3 Limitation of the Problems 5
1.4 Formulation of the Problems 5
1.5 Objectives of the Research 6
1.6 Uses of the Research 7

II. LITERATURE REVIEW

2.1 Review of Previous Research 8
2.2 Review of Related Literature 11
- 2.2.1 Concept of Test 12
- 2.2.2 Purpose of Test 12
- 2.2.3 Type of Test 14
- 2.2.4 Item Test Types 16
- 2.2.5 Item Analysis 21

2.3 The Criteria of a Good Test 22
- 2.3.1 Validity 23
- 2.3.2 Reliability 25
- 2.3.3 Discrimination Power 27
- 2.3.4 Level of Difficulty 29
- 2.3.5 The Quality of Options 30

III. RESEARCH DESIGN

3.1 Research Design.. 33
3.2 Setting of the Research... 33
3.3 Research Participants 34
3.4 Data Collecting Technique... 34
3.5 Data Analysis... 35
3.6 Research Procedure.. 36

IV. RESULTS AND DISCUSSIONS

4.1 Result of the Research... 43
4.1.1 Result of Validity 44
4.1.1.1 Result of Face Validity... 44
4.1.1.2 Result of Content Validity.. 46
4.1.1.3 Result of Construct Validity…………………………....... 51
4.1.2 Result of Reliability………………………………….. 56
4.1.3 Result of Discrimination Power…………………………...... 57
4.1.4 Result of Level of Difficulty………………………….......... 58
4.1.5 Result of the Quality of Option……………………………...... 59
4.2 Discussion…………………………………………………….... 108

V. CONCLUSIONS AND SUGGESTIONS

5.1 Conclusions 118
5.2 Suggestions 119

BIBLIOGRAPHY 121
APPENDICES 123
APPENDICES ITEMAN.. 141

LIST OF TABLES

Table **Page**

1. Analysis of the questionnaire of face validity.................................... 125
2. Analysis content validity of English semester test items.................... 127
3. Analysis of the questionnaire of construct validity............................. 129
4. Analysis reliability of English semester test items 132
5. Analysis upper group of English semester test items.......................... 131
6. Analysis lower group of English semester test items.......................... 132
7. Reliability of English semester test items 133
8. Analysis level of difficulty and discrimination power........................ 135
9. Analysis the quality of option alternatives based on ITEMAN 136

LIST OF APPENDICES

Appendices **Page**

1. Research Schedule 124
2. Analysis of the questionnaire of face validity.................................... 125
3. Analysis content validity of English semester test items.................... 127
4. Analysis of the questionnaire of construct validity............................. 129
5. Analysis reliability of Englih semester test item................................ 130
6. Analysis upper group of English semester test items......................... 131
7. Analysis lower group of English semester test items......................... 132
8. Reliability of English semester test items……………………… 133
9. The calculation of reliability……………………………………….. 134
10. Analysis level of difficulty and discrimination power………… 135
11. Analysis the quality of options based on ITEMAN………………..….. 136

I. INTRODUCTION

The first chapter deals with the background of the problems, the reason why the writer chooses the topic and supporting statement about it, identification of the problems, limitation of the problems, formulation of the problems and objectives of this research.

1.1 Background of the Problems

Testing refers to an effort to measure the result of student’s learning in teaching learning process. Consequently, the teachers should have an ability to arrange a good test and analyze of a test. Therefore, the accuracy and the carefulness of teachers may have a big impact on the increase of the quality of teaching particularly in giving the judgement of student’s ability. According to one of the expert that testing and teaching are closely interrelated those are virtually impossible to work in either field without being constantly concerned with the other. In other words, it is clear that teaching ought to be followed by testing. Without testing, it is impossible to evaluate and to measure the learning outcomes.

The arrangement of the test is a very important process because the teacher and other involved people are able to get information based on the test. A good test is important to measure how much students understand the material, and to determine the student give attentive to any material provided by the teacher in the learning process. The ability to formulate a good test that are needed by the teacher to evaluate, whether the instrument used in accordance with the desired, among others, it can determine the students who have mastered the material that is taught and the teacher can help to improve the test through revision or dispose of ineffective tests. The instrument is test items. A good and bad of the test can be viewed from several aspects, that is : Validity, a test can be considered to be valid if it can be measure what it is supposed to measure. Reliability, a test can be considered to be reliable if it can show a test consistent result. Discrimination power, a good test based on discrimination power is one which is able to differentiate between the upper group and the lower group. Level of difficulty, a good test based on level of difficulty is the test which is not too easy or too difficult.

When the writer observed some students in pre-research, the writer found the student’s difficulties in answering English semester test items. When the writer asked the students about their problems, part of them answered that English semester test items were too difficult to answer, and another students also said that sometimes English semester test items were difficult to answer because they have not learned them before. Then, there were some of students who said that sometimes the teacher’s explanation was not clear, therefore they did not understand the material, even there were some students who said that their difficulties in answering English semester test were because of their lack of interest in learning English. They learn English but they do not feel interested in it. It caused they are lack of awareness of how important to learn English. Besides, they are also lack of motivating factors to learn English.

It was also stated by English teachers at SMK Negeri 1 Gedong Tataan when interviewed by the writer in the pre-research, most of students in SMK Negeri 1 Gedong Tataan still got difficulty to answer the test items. The material being tested might not match the capabilities of students in these schools because English test items were not made by the school's teachers but by the teacher team of district called teacher team made test or MGMP (*Musyawarah Guru Mata Pelajaran*) Gedong Tataan, Pesawaran. Actually, testing is aimed to determine the achievement of the objective of education. Teacher as a constructor of the test should construct a good test so that the test will be valid and reliable. Test that is made by the teacher team, it is still to be questioned whether the test is valid and reliable or not because the teachers rarely tried out the test first before giving it to the students. Because a good test, without tryout is impossible. Knowing this fact, the teachers should analyze the the test so that the teacher will determine the quality of the test. By analyzing the test, the teacher will determine which items can be used or which items should be revised. And for the students, they are able to measure their ability in mastering the materials.

So, this information is very useful for both students in their learning and the teachers in their teaching. It can be a feedback for the teachers, who have responsibility to meet the instructional objectives, while for the students, it illustrates their performance. Related to the importance of the evaluation, it is necessary to consider that the test should be well constructed. As a means of evaluation, a test is administered to get information about the student’s improvement and to measure the result of the teaching learning process. And semester test is a test activity which is held at the end of teaching learning process in one semester. That is why, the writer assumes that semester test is one kind of tests which is intended as a feedback from the students and also as a result of teaching from the teachers in one semester.

Based on the explanation above, semester test is a tool of measurement to give some information related to student's ability. Then, this information will be used to consider and to decide several rules not only for the student's but also for the teachers in increasing the quality of teaching learning process. And the English test in Gedong Tataan is made by MGMP (*Musyawarah Guru Mata Pelajaran*). While MGMP itself consists of a team who has responsibility to design a test for each subject, it means that the semester test items are rarely analyzed by the teachers after they are tested. Related to the previous condition, it is needs to investigate and to describe the characteristics of English test mention above.

Therefore, the writer is interested in analyzing the semester test items which are tested for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year. The problem concerning with the analysis of test items is very important to be investigated because all the teachers who teach English in SMK Negeri 1 Gedong Tataan should be able to tryout all the test items that they made or the teacher team made before they use them to test the students. That is, all the items of the test should have good quality.

To achieve the goals above, the writer carried out the current research by analyzing the English semester test items at SMK Negeri 1 Gedong Tataan in 2012/2013 academic year based on the criteria of a good test. A good test should have (1) Validity, (2) Reliability, (3) Level of difficulty, (4) Discrimination Power, and (5) The Quality of Options. This research was concerned with the whole of test items designed by MGMP Gedong Tataan, Pesawaran. This includes test analysis and item analysis. The puprpose of test analysis is administered to determine and describe such criteria as face validity, content validity, construct validity, and reliability. And the item analysis is to determine which items are good and which items are bad in term of the level of difficulty, discrimination power, and the quality of options.

1.2 Identification of the problems

In relation to background of the problems above, the following problems can be found:

1. The English teachers are rarely tried out and analyzed the test to determine the quality of test items.
2. The quality of English semester test made by MGMP is not identified in term of validity, reliability, discrimination power, level of difficulty, and the quality of options..

1.3 Limitation of the Problems

Based on the identification of the problems above, the writer focused the research on the quality of test items used in semester exams.

1.4 Formulation of the Problems

Based on the limitation of the problem, the writer formulated the problem as follow: “Do the English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year fulfill the criteria of a good test. The criteria are specified below:

1. How is the validity of English semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year?
2. How is reliability of English semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year?
3. How is the level of difficulty of English semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year?
4. How is the discrimination power of English semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year?
5. How is the quality of options of English semester test items at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year?

1.5 Objectives of the Research

In relation to the research problems above, the objectives of this research was to determine the quality of English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year, especially in relation to:

1. Validity,
2. Reliability,

3. Level of difficulty,
4. Discrimination power, and
5. Quality of the options.

1.6 Uses of the Research

The findings of this research are expected to be beneficial for theoretical and practical developments.

a. Theoretically

- The results of the reseach are expected to support the existing theory, especially on the theory of assesment and evaluation as it will be discussed in Chapter 2.
- To be a reference for the future research.

b. Practically

- As the information for the English teachers and the readers about face validity, content validity, construct validity, reliability, discrimination power, level of difficulty, and the quality of options of English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year
- As a feedback for the English teachers about the criteria of a good test concerning semester test items.
- As a consideration for the English teachers whether the test needs to be revise or not, especially for preparation in facing the second semester.

II. LITERATURE REVIEW

This chapter discusses two major points: review of previous research and review of related literature.

2.1 Review of Previous Research

This section reviews several studies that have been conducted in relation to the quality of English test items in general. There are three researchers who have conducted research on the quality of English test items (Hayatunnisa 2003; Lestari 2010; and Putri 2009).

Hayatunnisa (2003) at SMA Al-Kautsar Bandar Lampung found that first semester English test based on the criteria of a good test for the second year students of SMA Al-Kautsar Bandar Lampung in the year of 2002/2003 is good. The research employed descriptive method. The data was collected through questionnare and document. The test paper, together with the questionnare, was analyzed to investigate face validity and construct validity. Then, the test paper together with the material in Guidelines for Teaching Program 1998 for the second year studets of SMA, was analized to examine content validity. The students' answer sheet numbering 48 were divided into two groups, 24 for upper students and 24 for lower students. The students' answer sheet were analized to investigate reliability, level of difficulty and discrimination power.

The test has good face validity for the reason that the respon that the respondent gave 61.06% yes-answer for the questionnare given. It implies that the respondent understand about the instruction of the test. The test has very good content validity since fifty-three items or 88.3% in the test described the subject matter and objective of teaching English available in GBPP 1998, but there were seven items or 11.6% which was lack of content validity since the material do not represent the material provided in GBPP 1998, has good consruct validity for the reason that respondent gave 77.8% for the questionnare given and it means that most items were in line with the theory of language. The items have a high reliability (r= 0.68), it implies that the test has a high consistency whenever it is tested. From the finding it can be seen that the test has a diverse discrimination power of the objective test: 31 items were considered poor, 26 items were considered satisfactory, and 3 items were considered good. The test has average level of difficulty as follows: 14 items were easy, 41 items were average and 5 items were difficult.

Another research was done by Lestari (2010) from Sebelas Maret University. The motive of the research was the existing phenomenon in the teaching and learning process which emphasizes its measurement through tests. The concern of the study was the appropriateness of multiple choice and essay test items. The study focused on the description of the test items' appropriateness based on the quantitative data. The subject of the study is the English final test items for the second semester of twelfth grade students of SMA Negeri 5 Surakarta in 2008/2009 academic year. The data were taken from 100 students in four classes. The appropriateness of the test items analyzed by using item analysis technique.

The analysis comprises three aspects, namely index of discriminating power, level of difficulty, and the effectiveness of distracters. The appropriateness of the three aspects must be fulfilled if the test item is multiple choice. The study results a description of each test item based on quantitative data proceeded in the item analysis. Global result showed that there were only 27.5% of the total test items in the type of multiple choice that fulfil criteria of a good test items analyzed from the three aspects. Meanwhile, the essay test items was satisfactory, and able to fulfill two criteria.

And another research also was done by Putri (2009) from Universitas Negeri Semarang. The research was conducted to analyze the test-instrument after being used for evaluation, to know whether or not the instrument was good for assessing the students' mastery. Moreover, the data from the test result were analyzed to determine whether or not the test appropriately match with the instructional objective or standard competence stated in the curriculum and to determine the item analysis including difficulty level, discrimination power, validity, and reliability. It was a quantitative study. In writing this thesis, the writer was conducted to field research to collect the data.

The test papers and students' work sheets were used to collect the data. Samples were taken practically by the use of random sampling. The data was established by using some procedures. The test papers consist of 50 items in the form of multiple choices. The students answer sheets are needs for analysis to find out the quality of the items based on item analysis. They were analyzed by using analysis procedures. Furthermore, the result of the analysis of this test tells that the questions of the test are related to the 2006 curriculum, but the topics of the questions were not related to the students' study program. In this final test, it was clear that this test is not valid and need some revisions.

To sum up, based on the previous studies, it can be stated that all the above mentioned studies reconfirmed the importance of analyzing the Englist tests. The studies had been carried out to investigate and to describe the criteria of a good test, these studies helped the researchers to understand how to measure the result of student's learning. Not only that, these studies focused on the quality of Englist tests. Besides adapting questions to investigate the result of student's learning, these studies can help the researchers to build their idea on how to analyze the English test based on the criteria of a good test.

However, there were still, at least, one issue that has not been found, that was an analysis of English semester test items based on the criteria of a good test. Therefore, this research was carried out to deal with that issue and to find out the quality of English semester test items for the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year whether or not fulfilled the following criteria of a good test: (1) face validity, (2) content validity, (3) construct validity, (4) reliability, (5) discrimination power, (6) level of difficulty, and (7) the quality of options.

2.2 Review of Related Literature

For the further explanation about the students' difficulties in semester test items, the writer explained some related literatures about test and the criteria of a good test.

2.2.1 Concept of Test

A test is used to see whether or not the test actually tests what should be tested. Tuckman (1995:8) defines *a test* as the process of assessing an activity, the process of activity and outcomes of a program for the objectives or the criteria determined. It means that *a test* a process that must be done in teaching learning activity.

Tinambunan (1988:3) defines *a test* an instrument or a systematic procedure for measuring a sample behavior. Shohamy (1985:3) supports that *a test* is a sample of knowledge and needs to be a good representation of it. It means that, what should be tested just a sample of behavior or knowledge, not the whole or behavior what the teachers has taught and the students have learned because it is also impossible to measure all of the students' abilities. The things that should be taken into account is the sample must be representative in the sense which is tested, it should reflect the knowledge that has been taught.

From the test are able to perform either in comparison with other or in comparison with a domain of performance task. Thus, a test is an instrument used to measure instructional objective or sample behavior.

2.2.2 Purpose of Test

Tests are used for many purposes. Shohamy (1985:6) mentions the purposes of classroom test are follows:

1. To find out whether what was taught was also successfully acquired
2. To evaluate and improve instruction
3. To obtain information on student's progress and language knowledge

4. To help organize learning/teaching material
5. To provide information to students on their language progress
6. To provide information for grades
7. To help diagnose students’ strengths and weaknesses in the language
8. To motivate students to learn

It can be said that the test is useful not only for the teachers but also for the students. For the teachers, it gives some information whether the material that was taught meets the objective that has been determined. It also can be used to measure the students’ progress whether they are success or failure in learning. Thus, the teachers can improve and evaluate either instruction or learning or teaching material to be better. While for the students, the test provides information on their progress. When the students determine what to expect on a test, they must study accordingly. Hence, the test would be able to motivate the students to learn. In addition, Austin as quoted by Hayatunnisa (2003:17) mentions that there were two features appering to be common to all testing programs.

1. Test results are used to make decision about the educational future or individual students.
2. The programme incorporate some methods of deciding which students have been succed and which students have failed.

So, there were two characteristics that evaluation not only provide information about the students’ achievement but also contribute the improvement of educational which involves the teachers, students, resources, methods and techniques of teaching directly, most of this information were collecting in the classroom through the test, and the way to arrange and analyze the test items.

2.2.3 Type of Test

Harrison as quoted by Hayatunnisa (2003:7) categorizes test accordingly to the need of the test and the use of the result of the test. Those four types are: *(a) Proficiency Test, (b) Achievement Test, (c) Diagnostic Test, and (d) Placement Test.*

a. Proficiency Test

Proficiency test (Harrison as quoted by Hayatunnisa, 2003:8) is designed to evaluate person's language knowledge in relation to the future language use. It does not necessary depend on what has been learned before in a given course, because it is concerned with the students' current knowledge in relation to their future needs. *Proficiency test* is the most suitable tool for assessing English for specific purposes. That is why, this test is usually used for looking for a job or continuing study. If it is for continuing study to foreign country, of course this test will cover requirement for being foreign language learner.

b. Achievement Test

Achievement test (Harrison as quoted by Hayatunnisa, 2003:8) tries to evaluate the test takers' language in relation to a given curriculum or material which the test-taker had gone through in a given course. It is intended to show the standard which the students have reached in relation to other students at the same stage. *Achievement test* covers a wider range material than a diagnostic test and relates to a long-term rather than short-term objective. For example, give the test to find out how much the students had learned.

c. Diagnostic Test

This type of test is used to identify the test taker's strength and weakness in the particular element of language as well as to attempt explaining why certain problems occur, and what treatment can be assigned (Harrison as quoted by Hayatunnisa, 2003:8). This kind of test can be conducted, for example at the end of the unit in the course book. In short, diagnostic test tries to provide the information about how well the students have learnt on the particular material.

d. Placement Test

Placement test (Harrison as quoted by Hayatunnisa, 2003:8) is designed to short new students into teaching groups, so that they can start a course at approximately the same level as the other students in the class. Thus, it relates to general ability of the test takers rather than specific points of learning.

While in this research, the test has been analyzed was *achievement test*, in the case of semester test, which was designed by MGMP Gedong Tataan, Pesawaran. Achievement test tried to investigate the students' achievement based on the objective of a given material.

2.2.4 Item Test Types

The item type which was used in classroom test was typically divided into two general categories (Grounlund, 2000:235), namely: (1) *the objective test items* and (2) *the subjective test items*. It means that, the first was *the objective test items* in which the answer decide right or wrong based on the key answer have been made, for example matching test items or multiple choice. The second was *subjective test items*. Scoring system of this type of test item was commonly based on the weight of item test, level of difficulty and how to close the students' answer some ideal answer, for example essay test. Subjective test items require students to write band present original answer. In short, *the objective test items* has only one correct answer per item, yet the subjective test items may result in a range of possible answer, some of which are more acceptable than other. *Here are some examples of objective test*:

a. Matching Test

Matching test consists of two parallel columns with each word, or symbol in one column be match to a word (Grounlund, 2000:235). The items in the column for which a match is sought, it is calling premise. While the item in the column from which the selection is made, it is calling response. *Matching test* is the most suitable for obtaining information about person knowledge of fact. It easy to be administered and to be scored. In contrast, *matching tets* is restricted to measurement of factual information base on learning and it is highly suspectible to the presence of irrelevant clues.

The matching response is self-evident but it is better to explain in the direction. In matching test, there are more names in column B are needs to match each event in column A. Then, the direction indicates that an item may be used only once, more than once, or not at all. Here is the example of matching test according to Grounlund (2000:236):

Table 1. Match each of words in A with its definition in B

A		*B*	
1	*The first US astronaut to ride in an space capsule*	*A*	*Edwin Aldrin*
2	*The first US astronaut to orbit in the earth*	*B*	*Neil Amstrong*
3	*The first US astronaut to walk in a space*	*C*	*Frank Bowman*
4	*The first US astronaut to step on the moon*	*D*	*Scot Carpenter*
		E	*John Glen*
		F	*Wally Schira*
		G	*Alan Shepard*
		H	*Edward White*

b. Multiple Choice

Multiple choice consists of a problem and a list of suggested solution (Grounlund, 2000:237). The problem may be stated in form of direct question or an incomplete statement and it is called *stem*. The lists of suggested solution are called *alternative*. It is recommended by using four alternative for grammar items, but five for reading and vocabulary. The correct alternative in each item is called *answer* and the remain alternative is called *distractors.*

In *multiple choice*, if there is one absolutely correct answer and all other alternatives are clearly wrong; this is known as the correct-answer type of multiple choice items. While the best-answer type of multiple choice is useful for measuring and learning outcomes that require understanding, application, or interpretation of factual information. These types of test are difficult to be constructed and give the students a possibility to guess, on the hand, it is easy to score, objective and reliable. Here are some suggestions before constructing Multiple Choice items according to Shohamy (1985:39).

A good suggestion to construct *multiple choice items* are started from open ended questions. Then make sure that the distractors are similar in terms of level of language, style of the language and the length. It is suggested to avoid items which can be answered without reading the text, just from general knowledge. General knowledge is background information that very helpful in reading as well as in testing reading. Make sure that there is as little possible as possible inter-dependency of items. That is, the students can answer one question based on the others. And the last is distractors should be logical continuation of the *stem* when the *stem* requires completion of the sentence. Here is the example of multiple choice tests according to Shohamy (1985:39):

What does the word "cautious" in line 4 mean?

a) *Careful*
b) *Famous*
c) *Lucky*

c. True/False Test

True/False test requires the test-taker to select the right answer out of two possible ones (Shohamy, 1985:41). And Heaton (1991:113) adds that the

True/False is the most widely used test reading comprehension. It means that the students have to decide whether it is true or false. This type of test is most suitable for reaching out the information about simple learning outcomes. *True/False test* is difficult enough to construct, because the distractors must be incorrect, but very efficient in scoring. In true/false, there are two main disadvantages: firstly, it can encourage guessing and secondly, the test may fail to discriminate widely among students.

Talking about guessing, there are two solutions that can be used to minimize it. The first, the correct answer will be awarded two, while for each wrong answer will be deducted from the score. So the students will not guess blindly. The second, the test include the third question in addition to True False option, for example True, False, not stated.

Table 2. The example of True/False test according to Shohamy (1985:42);

TRUE or FALSE ITEMS		*T*	*F*
1	*The fly into flew into the old man's eye*		
2	*The monkey hit the fly*		
3	*The monkey loved the old man*		
4	*The old man was working in the garden*		

d. Fill in the Blank

Cloze test is originally intended to measure the reading difficult level of a text (Heaton, 1991:131). In *close test*, the words are deleted systematically. So, once the actual text has been chosen, the construction of a close test is purely mechanical. The close tests are easy to construct and to score especially when it is used to exact scoring procedure. On the other hand, close test has a practice effect; the students who have practice doing it perform better on it. Close test is not really clear what it measures. Here is the example of C-Test according to Heaton (1991:132):

..... I have left you _____ which will make you _____ he told them. But _____ must dig in all _____ fields to find the _____ where the treasure is _____.

e. Open Ended Question

Open ended question test requires that the students answer the question by using their own words in written or oral, with no distractors. Actually, it is not too difficult in constructing the item but it is really difficult to score the answer, especially when it is a longer essay question. Then, the example of Open Ended Question according to Shohamy (1985:44):

Relating to the text of "The Monkey and the Fly" on page

1. *What the meaning of the word "chased" in line 3?*
2. *What happened to the old man's nose?*

f. Summary

A summary is a test which requires the students give synopsis of a given content (Shohamy, 1985:45). It means that the student asks to make a synopsis in oral or written form of a context given. The teachers only ask the students to summarize the material by using their own word. In contrast, this test item is really inefficient in scoring and costly. So, it is recommended to use this type of test in a classroom test, since it is possible to test a number of skills, for example reading in a simultaneous way. Then the example of summary test according to Shohamy (1985:45) is as follow:

Read the article "Smoke in Home Equals 20 a Day for Children".

Write a short summary (one or two paragraphs) of this article. Or what are the main points of the article?

2.2.5 Item Analysis

Item analysis is a process which examines the students response to individual test items in order to assess the quality of those items and of the test as a whole. *Item analysis* is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, *item analysis* is valuable for increasing teacher' skills in test construction, and identifying specific areas of course content which need the students emphasis or clarity (Scorepak®, 2008). It means that the quality of the test as a whole was assessed by estimating its "internal consistency." The quality of individual items was assessed by comparing students' item responses to their total test scores. An item analysis involves many statistics that could provide useful information for improving the quality and accuracy of multiple-choice or true/false items (questions).

The result of item analysis could be used to select items of desired difficulty that the best discriminate between high and low achieving students (Groundlund, 2000:315). It means that the results of an item analysis could be useful in identifying faulty items and can provide information about the students misconceptions and topics that need additional work. And Groundlund (2000:315) mentions the importance of item analysis. There were:

a. Item analysis data provide a basis for efficient class discussion of the test results.

b. Item analysis data provide a basis for remedial work.

c. Item analysis data provide a basis for the general improvement of classroom instruction.

d. Item analysis procedures provide a basis for increased skill in test construction.

While Anthony (1983:284) states that the importance of item analysis are determining whether an item functions as the teacher intends, feedback to students about their performance and as basis for class discussion, feedback to the teachers about pupil difficulties, areas for curriculum improvement, revising the items, improving item writing skills. Based on the explanation, the item analysis would be used to determine the level of difficulty, discrimination power, and option analysis.

2.3 The Criteria of a Good Test

A good test should fulfill certain the criteria. There are four criteria of a good test according to some expert; they are validity, reliability, level of difficulty, and discrimination power. Concerning about the criteria of a good test above, the writer was focused on the opinions. There were five components that the writer was going to research such as:

1) Validity

2) Reliability

3) Level of Difficulty

4) Discrimination Power

5) The Quality of Options

In relation to analyze of item English semester test items at the first year for the first semester of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year.

2.3.1 Validity

Validity refers to the extent to which an instrument really measures the objective to be measured and suitable with the criteria (Hatch and Farhady, 1982:250). A test can be considered to be valid if it can precisely measure the quality of test. It was also aimed to make sure whether the test has a good validity or not. This seems simple enough. In other words, a test can be said to be valid to the extent that it measures what it is supposed to measure. If the test is not valid for the purpose for which its design, the scores do not mean what they are supposed to mean.

1. Concept of Validity

Shohamy (1985:74) said that *validity* refers to the extent to which the test measures what it is intended to measure. It means that it relates directly to the purpose of the test. For example, if the purpose of the test was to provide the teachers with information whether the students could be accepted to a certain program, the test would be valid if the results given an accurate indication of that.

2. Types of Validity

Concept of validity reveals a number of aspects in which they drive several types of validity and attempt to show its relevance for the solution of language testing problem. To measure whether the test has a good validity, the writer was used face validity, content validity, and construct validity.

a. Face Validity

According to Heaton (1991:159), *face validity* concern with what the teachers and the students think of the test. It implies that face validity related to the test performance, how its look like a good test. Face validity only concerns with the layout of the test. Considering the importance of face validity, it was important to ask the teachers and the students to give their opinion about the test performance. In a formal way, face validity could be analyzed by distributing questionnaire. If a test does not appear to be valid to the students, they may not do their best.

b. Content Validity

Content validity was concerned with whether the test is sufficiently representative and comprehensive for the test. Shohamy (1985:74) defines that the most important validity for the classroom teacher is *content validity* since this means that the test is a good reflection of what has been taught and of the knowledge with the teachers want the students to know. *Content validity* is the most important aspect of validity because it also gave the information whether or not the students understand the material given.

It means that the items of the test should present the material being discussed. Then, the test was determined to based o the materials that have been taught to the students. In other words, the test was based on the materials in the English curriculum, so that it can be said that the test has content validity since the test was good representation of material studied in the classroom. Shohamy also adds that content validity can be best examined by the table specification. It was necessary for the teachers to make spesification list to ensure that the test reflects all areas to be assessed properly and to represent a balanced sample.

c. Construct Validity

A test can be considered to be valid if the item of the test measures every aspect which is suitable with the specific objective of the instruction. *Construct validity* would be concerned with whether the test is actually in line with the theory of what it means to know the language (Shohamy, 1985:74). It means that the test items should really test the students or the test items should really measure the students' ability in English semester test items. For example, if the teachers want to test about reading, the teachers have to be sure that the test item really tests about reading, no others. Thus, a test can be said to be construct valid if it measures the construct or theoretical ideas.

2.3.2 Reliability

Reliability refers to the consistency of measurement that is, to see how consistent test scores or other evaluation results are from one measurement to another (Grounlund, 2000:193). While Hatch and Farhady (1982:243) adds that reliability of a test can be defined as the extent to which a test procedure consistent result when administrated under similar condition. From those two opinions, if a test is administered to the same condition on different occasion, the extent that it produces different result, it is not reliable. Since reliability is a necesssary characteristics of any good test, so it is needs to keep the test reliable. According to Heaton (1991:169), there are some ways to keep the test reliable:

1. Increasing the sample of material select for testing. The larger the sample, the greater the probability that the test as a whole is reliable
2. Administration and scoring of the test. It is suggested to make a rating scale. So that the maker can identify precisely what he or she expects for each scale and assign the most appropriate grade to the task being assessed.

There are some methods that can be established in computing reliability according to Henning (1987:81): they are (1) *Test-Retest*, (2) *Parallel Forms*, (3) *Inter-Rater*, (4) *Split Half*, and (5) *Kuder Richardson* (*KR-20 and KR-21*). In this research, the writer assessed the reliability of the test by using formulation Kuder Richardson 21. But firstly, the writer calculated the total scores divided by the number of subject to obtain the mean. The writer used standard deviation. The purpose of obtaining the standard deviation is to measure the standard from the mean. It means that how get the individual data in a data set was dispersed from the mean. The formula of standard deviation according to Henning (1987:40) is:

$$S = \sqrt{\frac{\sum(X-x)^2}{N}}$$

Where,

S : the standard deviation

X : the student's score

x : the mean of value

N : the number of students

After getting the standard deviation, then the writer used *kuder richardson 21* formula (Henning, 1987:84) to determine the reliability of the whole test as follow:

$$Rt(KR21)=\frac{N}{N-1}\left(1-\frac{x(N-x)}{NS^2}\right)$$

Where,

N : the number of items in the test
x : the mean of the test scores
S^2 : the variance of the test scores
Rt : reliability

Tuckman (1995:256) states that the reliability of a test can vary between 0.00 and 1.00. A reliability of 0.00 indicates that a test has no reliability and hence is an inadequate test for making any judgement about the students. A reliability of 1.00 is a perfect reliability, indicating a perfect or error-free test. Reliability here is reported with numbers between 0.00 and 1.00. For computing the reliability of the test, the writer utilized *kuder richardson 21*, since it was simple enough. It just required three types of information, they were, the number of items, mean, and standard deviation of a test. And the correlation of coefficient would be interpreted by using the following criteria:

0.90 – 1.00 : High
0.50 – 0.89 : Moderate
0.00 - 0.49 : Low
(Hatch and Farhady: 1982:247)

2.3.3 Discrimination Power

Discrimination power is an aspect of item analysis, discrimination power tells about which is the item discriminates between the upper group students and the lower group students. Shohamy (1985:81) states that discrimination index tells about the extent to which the item differentiates between high and low students on that test. A good item based on discrimination power is one which is able to differentiate between the upper group and the lower group. It means that one which good students did well on, and bad students fail. If all the students answer a test item correctly at the end of instruction, this might indicate that both the instruction and the item have been effective.

Estimating discrimination power is subtracting the number of the lower group students who answer the item correctly from the number of the upper group students who answer the item correctly, and then divide the results by half number of both group students. The formula used to estimate discrimination power is as follows:

$$DP = \frac{U-L}{{}^{1}/_{2}T}$$

Where,

DP : Discrimination power
U : Upper group
L : Lower group
T : The total number of students
(Shohamy, 1985:81)

In accordance with Shohamy (1985:82), there are some criteria of discrimination power of an item. An item is excellent if the doscrimination index ranges from 0.71 to 1.00. a good item ranges from 0.41 to 0.70. A satisfactory item ranges from 0.21 to 0.40. An item is poor if the discrimination index ranges from 0.00 to 0.20, and an item is bad if the discrimination index is negative.

1. If the value is positive, it means that more high level students get correct answer than low students.
2. If the value is negative, it means that more low students than high level students got the item correct (it can be said that the test item is bad item, should be omitted).
3. If the value is zero, it means that there is no discrimination.
4. In general, the higher the discrimination index will be the better. In classroom situation most items should be higher than 0.20 indexes.

And Heaton (1991:180) adds the criteria of driscrimination power as follows:

DP : 0.00 – 0.20 is poor items

DP : 0.21 – 0.40 is satisfactory items

DP : 0.41 – 0.70 is good items

DP : 0.71 – 1.00 is excellent items

DP : Negative (Discarded, should be omitted)

2.3.4 Level of Difficulty

Difficulty level is one of kind of item analysis. Then, level difficulty was concerned with how difficulty or easy the item for the students. Shohamy (1985:79) states that difficulty level relates to how easy or difficult the item is from the point of view of the students who took the test. It is important since test items which are too easy (that all students get right) can tell us nothing about differences within the test population.

If the item too easy, it means that most or all of the students obtained the correct answer. In contrast, if the item is difficult, it means that most or all of the students get it wrong. Such item tells nothing about differences within the students. The difficulty level of item may range from 0.00 to 1.00 shows about the extent of difficulty level (Shohamy, 1985:73). Thus, if the difficulty level is 0.00, it means that the item is difficult. On the other hand, if the difficulty level is 1.00, it means that the item test is easy. The level of difficulty of objective test items is computing by using the following formula:

$$LD = \frac{U+L}{T}$$

Where,

LD : The level of difficulty

U : Upper group who got the item correct

L : Lower group who got the item correct

T : The total number of students

(Shohamy, 1985:79)

And the result was interpreted by looking at the following criteria:

LD : 0.00 – 0.30 is difficult

LD : 0.31 – 0.70 is average

LD : 0.71 – 1.00 is easy

(Shohamy, 1985:79)

2.3.4 The quality of options

Option analysis is a distribution of testees in diciding alternatives on a multiple choice test. It is obtained by calculating the number of testees who choose the alternatives A, B, C, or D or those who do not choose any alternatives. From this way, the teachers would be able to identify whether distracters function well or bad.

According to Suparman (1995:9) a distracter which is not chosen at all by all testees is a bad one because it is too outstanding. By contrast, a good distracter can attract the testees who do not master the content of the material well. A good distracter works well if it is chosen by at least 5% of testees and a distracter can be treated in one of the following three ways.

1. Accepting it, because it is goodc,
2. Rejecting it, because it is bad,
3. Revising it, because it is not so good. Its weakness may lie merely on its instruction.

To make easy distracter's analysis, the writer was used in ITEMAN. ITEMAN is one of the analysis programs that comprise Assessment Systems Corporation's Item and Test Analysis Package. ITEMAN analyzes test and survey item response data and provides conventional item analysis statistics (e.g., proportion/percentage endorsing and item-total correlations) for each item, in order to assist in determining the extent to which items are contributing to the reliability of a test and which response alternatives are functioning well for each item. And as additional data, to interpret the results of analysis item test, the writer used the criteria of quality test item by some experts in Suparman (2011).

Table 1. The criteria of test item quality

Criteria	Indeks	Classification
Prop. correct (p) (level of difficulty)	0.000 – 0.250	Difficult
	0.251 – 0.750	Average
	0.751 – 1.000	Easy

Point biser (D) (discrimination power)	D≤ 0.199	Very low
	0.200 – 0.299	Low
	0.300 – 0.399	Average
	D ≥ 0.400	High
Prop. Endorsing (the quality of options)	0.000 – 0.010	Low
	0.011 – 0.050	Sufficient
	0.051 – 1.000	Good
Alpha (reliability)	0.000 – 0.400	Low
	0.401 – 0.700	Average
	0.701 – 1.000	High

To make easier in choosing item test which need to be revised or dropped, recommended by using the following criteria:

Table 2. The criteria to classify the quality of the test items

Criteria	Indeks	Classification	Interpretation
Prop. correct (p) (level of difficulty)	0.000 – 0.099	Very difficult	Drop/needs total revising
	0.100 – 0.299	Difficult	Needs revising
	0.300 – 0.700	Average	Good
	0.701 – 0.900	Easy	Needs revising
	0.901 – 1.000	Very easy	Drop/needs total revising
Point biser (D) (discrimination power)	D≤ 0.199	Very low	Drop/needs total revising
	0.200 – 0.299	Low	Needs revising
	0.300 – 0.399	Quite average	Without revision
	D ≥ 0.400	High	Very good
Prop. Endorsing (the quality of options)	0.000 – 0.010	Least	Drop/needs revising
	0.011 – 0.050	Sufficient	Good enough
	0.051 – 1.000	Good	Very good
Alpha (Reliability)	0.000 – 0.400	Low	Not sufficient
	0.401 – 0.700	Average	Sufficient
	0.701 – 1.000	High	Good

III. RESEARCH DESIGN

This chapter discuses the design of this research and how to collect the data from the research participants. The writer also encloses the data collecting technique, the procedures of this research, the scoring system and how to analyze the data.

3.1 The Research Design

The design of the research was descriptive analystic. This research was intended to determine whether or not the first semester English test for the first year students of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year meets such criteria as face validity, content validity, construct validity, reliability, difficult level, discrimination power, and the quality of options. Descriptive analystic was a kind of method which is used to evaluate the document without reducing or adding. The data was authentic data, as it is.

3.2 Setting of the Research

1. **Time**

   The research conducted in a week. It would administered during the English lesson which is being tested when the students had been finished their English semester test items, and the writer asked their time to questionnaire about face validity.

2. **Place**

This research conducted at the first year of SMK Negeri 1 Gedong Tataan. There were 2 classes of first year class in the school, Computer Department and Automotive Department.

3.3 Research Participants

This research, the writer chose the first year students in the first semester of academic year 2012/2013 was observed. There were two classes of first years in the school, Computer Department and Automotive Department which consist of 24 up to 34 students in each class. Both of class used for research participants. To complete the data, the writer involved the English teachers and the experts as the second observe.

3.4 Data Collecting Techniques

The data were analyzed in mainly the test paper with the answer sheets. There were two techniques to collect the data:

1. **Document**

In getting the data, the writer used document. Document is one source of data in qualitative research. According to Setiyadi (2006:249) document has some strenghts compared with other data sources such as it can be obtained data easily and it can get the natural data. Document is scientific data that is not reactive while the subject can not hide anything. Document can be variety of forms, from the data are very personal up to very formal.

And in this research, the document was used to identify the English semester test papers and the answer keys of the English semester test that has been tested, and the answer sheets from the teacher as well as curriculum.

2. **Questionnaires**

The writer distributed questionnaires after conducting the observation. The writer also got more sources of the data from the subjects. According to Setiyadi (2006), there were two types of questionnaires, close-ended and open-ended questionnaire. In this research, the writer used close-ended questionnaire while the students have to choose the available answer in the questionnaire. Questionnaire was used to collect the data of face validity and construct validity of the test. The questionnaire of face validity gave to the first year students of SMK Negeri 1 Gedong Tataan since they have been taught by using the material which is tested in the achievement test. The questionnaire consist of 2 response category, *"Ya"* or *"Tidak"*. It consisted of 8 questions. The students were instructed to answer a set of question related to their thought about the layout of the test. And the other technique, the questionnaire of construct validity distributed to English teachers of SMK Negeri 1 Gedong Tataan because they were familiar with the material that was tested and the experts.

3.5 Data Analysis

In analyzing the data, the writer used test analysis and item analysis. Test analysis as examination to evaluate the students. Test analysis was intended to analyze the whole test for determining the quality of the test, such as face validity, content validity, construct validity, and reliability. While Item analysis was a process which examines the students' response to individual test items in order to assess the quality of those items and of the test as a whole. And item analysis were utilized for investigating such criteria as difficulty level, discrimination power, and the quality of options. In analyzing the quality of option alternatives, the writer was used to ITEMAN and also as supporting data.

3.6 Research Prosedure

The procedure of this research carried out in some steps in test analysis and item analysis:

a. The writer implements the steps of the test analysis as follows;

1. Found the research participants
2. Taking all the test paper that would be analyzed
3. Arranging all the answer sheets from the highest score to the lowest score
4. Determining face validity of English test, the questionnaire of face validity distributes to the students who take the test that would be chosen randomly. Then, the table below was used to check the face validity of the test by calculating yes-answer from the total answer.

Table 2. The Sum of the Questionnaire data

Respondents	The Item of Questionnaire								Sum	
	1	2	3	4	5	6	7	8	Yes	No
1										
2										
3										
4										
5										
6										
Sum										

Note: Y stands for yes-answer, N stands for no-answer.

Then, the data will be calculated as follows:

$$\frac{yes-answer}{total\ answer} \quad x\ 100\% = \ldots\ldots$$

The range of the result with its qualitative interpretation based on Arikunto's (1997:208) as follows:

Range	Qualitative Interpretation
81% - 100%	Very Good
61% - 80%	Good
41% - 60%	Fair
20% - 40%	Bad
Less than 20%	Very Bad

5. In determining the content validity, the writer analyzed the test items by comparing the test items with Syllabus for the first semester of the first year of SMK. The result presented on the table below:

Table 3. Analysis content validity of objective test items

No	Material in Curriculum			Test Items	Sum	Percentage
	Aspect of English being taught	Theme	Sub Theme			
Total						

6. In determining construct validity, the writer analyzed the test paper and questionnaire of construct validity to see whether or not the tests have construct validity. While the way to calculate content validity was the same as calculation in face validity. Below was the proportion of yes-answer from the total answer to see how well the construct validity of the test. And the data calculated as in the data of face validity.

Table 4. Analysis of construct validity

Respon-dents	The Items of Questionnaire								Sum	
	Listening		Vocabulary		Reading		Grammar			
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Sum										

7. In calculating reliability of the test, the writer used KR 21. The first steps was estimating standard deviation, the writer used the following formula:

$$S = \sqrt{\frac{\sum(X-x)^2}{N}}$$

Where,

s : the standard deviation

x : the student's score

$\bar{x}$: the mean of value

N : the number of students

The writer used KR 21 for computing the reliability of English semester test items.

$$\text{Rt(KR21)}=\frac{N}{N-1}\left(1-\frac{x(N-x)}{NS^2}\right)$$

Where,

N	: the number of items in the test
x	: the mean of the test scores
S^2	: the variance of the test scores
Rt	: reliability

As the last step, the correlation of coefficient was interpreted by using the following criteria:

0.90 – 1.00	: High
0.50 – 0.89	: Moderate
0.00 - 0.49	: Low

(Hatch and Farhady: 1982:247)

b. The writer implemets the steps of item analysis as follows:

The steps of items analysis were done as follows :

1. The first step was similar to that of the test analysis, arranging the entire answer sheets from the highest score to the lowest score.
2. Identifying an upper group separately by selecting the sheets with the highest for the upper group, and the sheets with the lowest score for the lower group.
3. Then, the writer calculated discrimination power (DP). The computation of discrimination power was held as follows:

   a. Subtract the number of student in lower group who got the item correct

   b. The result of the subtraction divided by the half number of two groups

   c. The formula of discrimination power was as follows:

   $$DP = \frac{U-L}{{}^{1}/_{2}\,T}$$

   Where,

   DP : Discrimination power
   U : Upper group
   L : Lower group
   T : The total number of students (upper and lower group)
   (Shohamy, 1985:81)

   d. As the last step, the result was interpreted by using the criteria as follows:

   DP : 0.00 – 0.20 is poor items

DP	: 0.21 – 0.40 is satisfactory items
DP	: 0.41 – 0.70 is good items
DP	: 0.71 – 1.00 is excellent items
DP	: Negative (Discarded, should be omitted)

(Heaton, 1975:180)

4. While for calculating the level of difficulty (LD) of each item, the following computation was used:
   a. Adds the number in the upper group who got the item correct to that in the lower group who got the item correct
   b. The result of the addition divided by the two groups. The formula for computing the level of difficulty as follows:

$$LD = \frac{U+L}{T}$$

Where,

LD	: The level of difficulty
U	: Upper group who got the item correct
L	: Lower group who got the item correct
T	: The total number of students

(Shohamy, 1985:79)

   c. At the last, the result was interpreted by using the following criteria:

LD	: 0.00 – 0.30 is difficult
LD	: 0.31 – 0.70 is average
LD	: 0.71 – 1.00 is easy

(Shohamy, 1985:79)

d. The result of the item analysis above presented on the following table:

Table 5. Level of Difficulty and Discrimination Power

Items	Key answer	UG	LG	U+L	U-L	DP		LD	
						Result	Interpretation	Result	Interpretation

5. To determine the quality of options alternatives from the students patterns' answer, the writer was used in ITEMAN and in the following table:

Table 6. The Analysis of the Quality of Options Alternatives

Items	Alt.	Prop. Endorsing			*Point biser*	Clasiffication
		Total	Low	High		

V. CONCLUSIONS AND SUGGESTIONS

This chapter is intended to draw conclusions and put forward some suggestions as will be address in the following sections:

5.1 Conclusions

Based on the results of the data analysis, the test has been analized was achievement test, in case of semester tests, which was designed by MGMP Gedong Tataan, Pesawaran. Achievement test investigated the students' achievement based on the objective of a given material and the writer draws the following conclusions:

1. The validity of English semester test items administered at the first semester of the first year of SMK Negeri 1 Gedong Tataan in 2012/2013 academic year was catagorized as good because the fifty objective test items represented on the material available in English Curriculum 2006.
2. The reliability of English semester test items was calculated using KR 21. The result shows that the reliability of English semester test items was low reliability (r = 0.07) and it means that the items on the test needs revised.
3. The discrimination power of English semester test items based on manual data of the test shown on Table 8 (see Appendix 9) were 10 poor items, 24 satisfactory items, and 12 good items. There was one excellent item on this achievement test and 3 items were negative discrimination.

4. The level of difficulty of English semester test items based on manual data of the test shown on Table 8 (see Appendix 9) were 17 easy items, 16 difficult items , and 17 average items. If it is seen from both DP and LD in the semester test items, there were 3 items that should be discarded since that were very poor. There were 25 items which should be revised and the number of semester test items that can be used was 22 items.
5. The quality of options of semester test items based on ITEMAN shown on Table 9 (see Appendix 10). There were 15 key answers needs revising because the other alternatives have good chance better than the key answers have been fixed. There were 29 alternatives have rejected, 60 alternatives accepted, and 111 alternatives revised from 200 option alternatives.

5.2 Suggestions

In line with the conclusions described in the section previous, the writer would like to give the following suggestions:

1. It would be better for MGMP who design the test to analyze it after administering to the students. It is done to determine whether the test has fulfilled the characteristics of a good test, that is, validity, reliability, discrimination power, level of difficulty, and the quality of options.
2. The teachers are suggested to receive a kind of feedback for the results of the analysis improvement of the test designing.
3. The teachers should be tried out the test first before giving it to the students.
4. The teachers should be trained on how to analyze the tests effectively and efficiently and to revise bad test items.

5. Nowadays, analyzing test items is easy, especially for the teacher because there are some software that can be used for analyzing, such as Software ITEMAN, Anates, Microsoft Excel, SPSS (Statistical Program for Social Science).

BIBLIOGRAPHY

Anthony J. 1983. *Educational Tests and Measurement an Introduction*. New York : Harcourt Brace Jovanovichi, inc.

ASC. 1989-2006. *User's Manual for the ITEMAN Conventional Item Analysis Program*. St. Paul, Minnesota: Assesment Systems Corporation.

Hatch, E, and H. 1982. *Research Design and Statistic for Applied Linguistic*. London: New Burry House, inc.

Hayatunnisa, 2003. *An Analysis of the First Semester English Test for the Second Year Students of SMU AL-KAUTSAR Bandar Lampung*. Unpulished Script. Bandar Lampung: FKIP University of Lampung.

Heaton, J.B. 1991. *Writing English Language Tests*. New York: Longman inc.

Henning, G. 1987. *A Guide to Language Testing*. Mass : New Bury House Publishers.

Lestari, A. 2010. *An Analysis on The English Final Test Items for The Second Semester of Twelfth Grade Students of SMA Negeri 5 Surakarta in 2008/2009 Academic Year (A Descriptive Study).* Thesis. Surakarta: English Department of Teacher Training and Education Faculty. Sebelas Maret University.

Ngadimun. 2004. *Analisis Butir Soal dengan Komputer dan Menafsirkannya*. Makalah disampaikan pada sosialisasi KBK bagi guru SMP Kabupaten Tanggamus di Pulau Panggung, tanggal 22-24 Juli 2004. Bandar Lampung: HEPI.

Nurkanca, W and Sumantara. 1995. *Evaluasi Pendidikan*. Surabaya: Usaha Nasional.

Putri, Y. 2009. *Analysis of Teacher-made English Final Second Semester Test for the Year Eleven Students of SMA N 1 Ambarawa in the Academic Year of 2008/2009 Based on the Representativeness of Content Standart* . Under Graduates thesis: Universitas Negeri Semarang.

Setyadi, B. 2006. *Metode Penelitian untuk Pengajaran Bahasa Asing*. Yogyakarta: Penerbit Graha Ilmu.

Shohamy, E. 1985. *A Practical Handbook in Language Testing For the Second Language Teacher*. Tel Aviv: Tel Aviv University.

Suparman, U. 1995. *Language Testing and Evaluation Volume 2 (Measurement)*. FKIP University of Lampung.

Suparman, U. 2011. *The Implementation of Iteman to Improve the Quality of English Test Items as a Foreign Language (An Assesment Analysis). Aksara Jurnal Bahasa, Seni, dan Pengajarannya.* Vol. XII. No.1. Hal. 1-96. ISSN 1411-2501. Bandar Lampung: Jurusan Pendidikan Bahasa dan Seni, FKIP University of Lampung.

Tinambunan, W. 1988. *Evaluation of Student Achievement*. Jakarta: Depdikbud.

Tuckman, B. 1975. *Measuring Educational Outcomes Fundamental of Testing*. New York Chicago San Fransisco Atlanta: Hourtcart Javanovich, inc.

Universitas Lampung. 2007. *Pedoman Penulisan Karya Ilmiah*. Lampung: Lampung University Press.