ANALYZING THE QUALITY OF THE FINAL SEMESTER TEST BY USING ITEMAN AT THE SECOND GRADE OF SMP NEGERI 2 SUMBEREJO TANGGAMUS IN 2016/2017 ACADEMIC YEAR

(A Script)

By
Nurul Amanah

ENGLISH EDUCATION STUDY PROGRAM
ARTS AND LANGUAGE EDUCATION DEPARTMENT
TEACHER TRAINING AND EDUCATION FACULTY
LAMPUNG UNIVERSITY
2017

ABSTRACT

ANALYZING THE QUALITY OF THE FINAL SEMESTER TEST BY USING ITEMAN AT THE SECOND GRADE OF SMP NEGERI 2 SUMBEREJO TANGGAMUS IN 2016/2017 ACADEMIC YEAR

By

Nurul Amanah

The objectives of this research are to identify the validity, the reliability, the level of difficulty, the discriminating power andthe quality of the alternatives of the final semester test at the second year of SMP Negeri 2 Sumberejo in 2016/2017 academic year. The researcher investigated the final semester test which consisted of 50 items. The result shows that the construct validity is valid, the content validity is valid, but the face validity is not valid. The reliability is 0.592 or in average level. The level of difficulty of the test which consists of 22 items (44%) isacceptable, 21 items (42%) need revising, and 7 items (14%) should be dropped. The discriminating power which consists of 15 items (30%) is acceptable, 12 items (24%) are need revising, and 23 items (4%) need dropping. The quality of the alternatives is 26 options (15%) acceptable, 89 options (85%) need revising and need dropping. In general, it shows thatthe quality of the test is moderate.

Keywords: quality of the alternatives, reliability, validity.

ANALYZING THE QUALITY OF THE FINAL SEMESTER TEST BY USING ITEMAN AT THE SECOND GRADE OF SMP NEGERI 2 SUMBEREJO TANGGAMUS IN 2016/2017 ACADEMIC YEAR

By

NURUL AMANAH

A Script

Submitted in a partial Fulfillment of
The Requirements for S-1 Degree
In
The Language and Arts Departement of
Teacher Training and Education Faculty

ENGLISH EDUCATION STUDY PROGRAM
ARTS AND LANGUAGE EDUCATION DEPARTMENT
TEACHER TRAINING AND EDUCATION FACULTY
LAMPUNG UNIVERSITY
2017

Research Title : **ANALYZING THE QUALITY OF THE FINAL SEMESTER TEST BY USING ITEMAN AT THE SECOND GRADE OF SMPN 2 SUMBEREJO IN 2016/2017 ACADEMIC YEAR**

Student's Name : *Nurul Amanah*

Student's Number : **1313042059**

Department : **Language and Arts Education**

Study Program : **English Education**

Faculty : **Teacher Training and Education**

APPROVED BY

Advisory Committee

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

UNIVERSITAS LAMPUNG

ADMITTED BY

1. Examination Committee

   Chairperson : **Drs. 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

KEMENTERIAN RISET, TEKNOLOGI DAN PENDIDIKAN TINGGI
UNIVERSITAS LAMPUNG
FAKULTAS KEGURUAN DAN ILMU PENDIDIKAN

Dr. H. Muhammad Fuad, M.Hum.
NIP 19590722 198603 1 003

Graduated on : **August 09th, 2017**

SURAT PERNYATAAN

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

Nama : Nurul Amanah
NPM : 1313042059
Judul skripsi : Analyzing the Quality of the Final Semester Test by using ITEMAN at the Second Grade of SMPN 2 Sumberejo in 20116/2017 Academic Year
Program studi : Pendidikan Bahasa Inggris
Jurusan : Pendidikan Bahasa dan Seni
Fakultas : Keguruan dan Ilmu Pendidikan

Dengan ini menyatakan bahwa

1. Karya tulis ini bukan saduran atau terjemahan, murni gagasan, rumusan, dan pelaksanaan penelitian/implementasi saya sendiri tanpa bantuan dari pihak manapun kecuali arahan pembimbing akademik dan narasumber di organisasi tempat 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 ketidak benaran 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, 15 Agustus 2017
Yang membuat peryataan,

METERAI TEMPEL
26DFIADC002842722
6000
ENAM RIBU RUPIAH

Nurul Amanah
NPM 1313042059

CURRICULUM VITAE

The researcher name is Nurul Amanah. She was born in March 8th, 1994 in Sumberejo, Tanggamus. She is the second daughter from Karnyoto and Siti Maryam.

Her educational background started at SDN 1 Sumberejo in 2000 and graduated in 2006. After that, she registered in SMPN 2 Sumberejo and graduated in 2009. Then, in 2009, she studied in SMAN 1Sumberejo and graduated in 2012.

She was accepted in English Education Study Program of Teacher Training and Education Faculty in Lampung University through PMPAP in 2013. She accomplished her KKN-KT in July to August 2016. She taugh televenth grade students of SMA Kesuma Bhakti, Bekri, Lampung Tengah. She had examination of her script on August, 2017.

DEDICATION

I proudly dedicate this script to:

My beloved parents, Karnyoto and Siti Maryam

My lovely brother, Hendri Saktiawan

My lovely sister, Galuh Tri Wahyuning Tyas

My almamater, Lampung University

And all the seekers of knowledge in the world

MOTTO

"Smile is simple way of enjoying life"
(anonymous)

"Believe yourself as nothing, trust your Allah as everything"
(anonymous)

ACKNOWLEDGEMENTS

Alhamdulillah, praise is only rendered to Allah SWT, the Almighty, for the mercy and strength so that the researcher can finish the script entitled *Analyzing the Quality of the Final Semester Test by using ITEMAN at the Second Grade of SMPN 2 Sumberejo in 2016/2017 Academic Year*. This script is submitted to fulfill one of the requirements for obtaining award of the S-1 Degree Program at the Department of Language and Arts of Teacher Training and Education Faculty, University of Lampung.

It is important to be known that the script would never have come into existence without any supports, encouragements and assistance by several great people.Here the writer would like to address her gratitude and respect to:

1. Drs. Ujang Suparman, M.A. Ph.D., her first advisor, for her willingness to give assistance, ideas, invaluable evaluations, comments, suggestions and encouragement within her time during the script.
2. Drs. Ramlan Ginting Suka, M.Pd., her second advisor for his guidance, important evaluations, comments, and suggestions which are very beneficial for the improvement of the script.
3. Prof. Dr. Cucu Sutarsyah, M.A., her examiner who generously contributes his suggestion and criticism for the improvement of this script.
4. All lectures of English Education Study Program in University of Lampung who guide and deliver knowledge to her.
5. Big family of SMPN 2 Sumberejo, Tanggamus for providing the opportunity to conduct this research, the students of Class VIII 3, for their willingness, cooperation, and participation in this research.
6. Karnyoto and Siti Maryam, her beloved parents, who always give their immeasurable love, endless prayers, patience to wait for her graduation as well as supports given to keep her spirit alive, and encouragements for her success.
7. Hendri Saktiawan and Galuh Tri Wahyuning Tyas, her lovely brother and sister, who always give support for her after her parents.
8. All English Department 2013 students who share the unforgetable moments.
9. Her best friends; Ade Aprilia, Endah Triasih Br. Sihaloho, Hesti Karlina,Citra Mustika Zulya and Nur Choironi,who always on her side in every situation.
10. The meritorious friends, Agung Putra Mangunsong, Ade EkaSafitri, Dewi Purnama, Inkholisatun Nafsiah, Rima Varadina, Uun Yukanah, Yuniska Murti Ayu, Elza Yulistiana, Temu Riyadi, Claudia Citra, Rizki Dilla Sintya, Nurfadillah, Era Puspita, Niddia Raisa Marta and Alien Elanda.

11. Adi Purwoko, herbest partner, for always encouraging, motivating, supporting and helping her in every situation when she faces the difficulties in completing this script.

Hopefully, this research would give positive contribution to educational development or those who want to carry out further research.

Bandar Lampung, 09 Agustus 2017

Nurul Amanah

CONTENTS

Page

REFERENCES

APPENDICES

TABLES

TABLES **Page**

APPENDICES

APPENDICES **Page**

CHAPTER 1

INTRODUCTION

This chapter concerns with several sub chapters, that is 1). Background of the problems, 2).Identification of the problems, 3).Limitation of the problems, 4). Formulation of the research questions, 5). Objectives, and 6). Significance of the researchas elaborated in the following sections.

1.1. Background of the Problems

ITEMAN is one of the way to analyze the quality of the test especially multiple choice test. But unfortunately, there are so many teachers do not familiar about ITEMAN. Most of the teachers still using manual way to analyze students ability by counting the scores one by one. This way consumes much times. By using ITEMAN only need one click after input the data, the result of our data will be seen.. It can reduce time in scoring students test.

Final semester test is the important thing in assessing students ability. By using test, teachers will know the teaching learning is successful or unsuccessful. According to Arifin (2011 : 118) “test is a technique or procedure which is used in order to implement the activity of measurement, in which there are many questions or assignments should be done by a student.” Arikunto (2009: 51) explain that test is a tool or a procedure that used to know the result of something, with procedure certain that has been decided. The other definitions about test is a gatherer instrument of information but compared with other tools, this test is more legitimate because full of restrictions (Daryanto, 2008: 25). According to Sanjaya (2008. 239) test is a measurement instrument that is used for measuring the success of student in reaching certain competency. Multiple choice test is one of the most famous test which is used to test the students knowledge.

Multiple choice testing is an efficient and effective way to measure students' ability and form an objective assessment in which respondents are asked to select the only correct answer out of the choices from a list. Multiple choice tests have several advantages, if the writers are well trained and items are quality assured, it can be a very effective assessment technique. If the multiple choice tests have good quality, it can be performed better for assessing the students. Multiple choice tests often require less time than would tests requiring written responses. Multiple choice questions lend themselves to the development of objective assessment items, but without author training, questions can be subjective in nature.

In English assessment by using ITEMAN software program, there are some aspects to determine whether the test items are good or not, such as the validity, reliability, level of difficulty, discriminating power, quality of key answer and distractor. By using ITEMAN the quality of the test will be known. ITEMAN helps us to make the qualified test items for English assessment. ITEMAN is very good for helping teacher and test maker to know the quality of their test items in particular multiple choice test.

However, multiple choice tests have drawbacks, such as, students can guess the answers, the test does not measure deep thinking skills, writing successful multiple choice questions is difficult, and the students cannot organize and express their ideas, this kind of test is still popular because they are truly reliable and objective. This standardized test is also practical. It means that the test is easy to administer and consume less time to be assessed versus other assessments.

Multiple choice tests are possible ambiguity in the examinee's interpret information as the test maker intended can result in an incorrect response, even if the taker's response is potentially valid. Even if the students have some knowledge of a question, they receive no credit for knowing that information if they select the wrong answer. Another disadvantage of multiple choice examinations is that the student who is answering a particular question can simply select a random answer and still have a chance of receiving a mark for it. If randomly guessing an answer, there is usually a 25 percent chance of getting it correct on a four answer choices question. It is common practice for students with no time left to give all remaining questions random answers in the hope that they get at least some of them right.

Considering the facts above, the researcher try to help the teachers determine the quality of multiple choice tests by using ITEMAN program. ITEMAN is very important for the teachers taking charge of administering test in order to be sure about the quality of the test they use. Consequently, understanding how to interpret and use information based on student test scores is as important as knowing how to construct a well-design test.

ITEMAN comes from “item and test analysis”. ITEMAN is a software used to analyze test item and determine which test item is good and which is not, based on the criteria of reliability, discriminating power, level of difficulty, quality of key answer and the quality of the distractors. The data are analyzed automatically by the software; therefore the researcher is made easier and faster to do the analysis. The researcher does not need to acquire complicated mathematic calculation, since the steps are very simple to follow. Not only that, the software program could be used to analyze almost unlimited number of testees in relatively very short time.

ITEMAN can be defined as one of the analysis programs that comprises assessment systems of test items and test analysis package, (Assessment Systems Corporation (ASC), 1989-2006). As ITEMAN is consider useful, the teachers are more expected to have an involvement in assessing the multiple choice tests using the item analysis program.ITEMAN isimportant especially for lecturers and teachers of English who are responsible for administering tests (such as mid semester and final semester examinations). In order to utilize the program, the ITEMAN software program should be installed first.

Basically, ITEMAN can be used to analyze test and survey item-response data and provide conventional item-analysis statistics (e.g., proportion/percentage endorsing and item-total correlations) for each items. Such function is very important for English lecturers and teachers at school levels in order to assist them 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. Besides item-level statistics, more importantly the ITEMAN program also provides statistical indicators on the performance of the test as a whole.

The good school should be has bank of the test. Bank of the test is the test that has been tried out. The teachers almost do not know whether the test they are made good or not. The test made should be tried out first, so ITEMAN make it easier and faster in analyzing the quality of the test after being tried out. The researcher uses ITEMAN software program which helps the teachers determine the quality of the final semester test and prove whether the test has fulfilled the criteria of a good test or not. If the tool of the assessing is good, the result of the assessing is good, but if the tool is bad, the result will be bad too.

1.2. Identification of the Problems

Based on the background of the problem, the researcher finds several problems that can be identified. The multiple choice tests:

1. it is reliability may not be reliable,
2. the item may be too difficult or too easy,
3. the item need dropping or total revising,
4. the proportion of the answers still low,
5. the distractors do not function well,
6. the item does not check on the knowledge of the subject taught before,
7. the item does not discriminate the more knowledgeable students from the less knowledgeable students,

8. the item does not measure deep thinking skills, and
9. the item promotes confusion to the students.

1.3. Limitation of the Problems

Considering the identification of the problems, this research isfocused on the following issues:

1. validity of the test,
2. reliability of the test,
3. level of difficulty,
4. discriminating power of the test, and
5. quality of the alternatives.

1.4. The Formulation of the Research Questions

In line with the limitation of the problems, the following research questionsare formulated, as follows:

a. How is validity of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus?
b. How is the reliability of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus?
c. How is the level of difficulty of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus?

d. How is the discriminating power of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus?

e. How is the quality of the alternatives of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus?

1.5. The Objectives

In line with the formulation of the research questions, the objectives of this research are:

to find out how the quality of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus is, in relation to:

a. find out how the validity of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus is,

b. find out how the reliability of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus is,

c. find out how the level of difficulty of the final semester test at second grade semester of SMPN 2 Sumberejo Tanggamus is,

d. findout how the discriminating power of the final semester test at the second grade of SMPN 2 Sumberejo Tanggamus is, and

e. findout how the quality of the alternatives of the final semester test at second grade of SMPN 2 Sumberejo Tanggamus is.

1.6. The Significance of the Research

The findings of the research are expected to be beneficial both theoretically and practically.

1. Theoretically, as a verification of the previous theories of the quality of assessment.
2. Practically, this research may be used to help teachers assess the quality of multiple choice tests by using ITEMAN program.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter provides two major points. That is, review of the previous research and review of related literature. They are elaborated in the following sections.

2.1. Review of Previous Research

In relation to this research, there is some previous research which has been conducted by some researchers, such as: Fitriana, 2013;Ratnaningsih, 2009; Bagusalghani, 2015 and Nurung, 2008.

Fitriana (2013) conducted research on students of MI Sultan Agung at grade five, Sleman district, Yogyakarta. She investigated the quality of final semester test. She used ITEMAN as the tool to determine the quality of the test. The result of the research revealed that the multiple choice test made by official government (Dikpora),Sleman district, had high validity. There were 27 questions or 67.5 % of the test which were valid and 13 questions or 32.5% were not valid. There were 67 alternatives out of 120 alternatives were functional. Not only the validity, the alpha of the test was 0.780 meaning that it had high reliability. From the level of difficulty, it showed that there were 25 questions which were easy. The discriminating power which was accepted was 22 questions, because only 37.5 % of the multiple choice tests had good discriminating power.

Ratnaningsih (2009) conducted research on students of UT Pondok Cabe, Pamulang, South Tangerang town. The paper aimed to analyze multiple choice items of the End Semester Examination of UT using the program ITEMAN. The data used were the answer sheets of students taking eight courses in the first and second semester of 2009.The results showed that the test items used had a pretty good quality. Average test item difficulties were fair. This was indicated by the mean value of which ranged from 0.328 to 0.461. Discrimination index for both semester tests were good in about 75% of the courses measured. Its value ranged from 0.304 to 0.451 for the first semester of 2009 tests and 0.343 to 0.382 for the second semester of 2009 tests. Meanwhile, the reliability of the test items could be considered good whose value ranged from 0.771-0.520. The effectiveness of the alternatives was 62% - 94%. It meant that the alternatives were functional.

Bagusalghani(2015) conducted research in SMAN 1 Purbolinggo. He found reliability of the final test is 0.448, the level of difficulty; good or directly useable 30%, very difficult or needs revising 10%, very easy or needs revising 20%, too difficult or needs dropping or total revision 40%, The discrimination power; high 25,7%, average/without revising 0%, very low/needs dropping 51,5%, negative discrimination 17,1%, The quality of the alternatives it was found that the alternative of the 35 items consisting of A, B, C, D, and E with the total of the alternatives is 175, can be classified into three categories, that is, *very good*, *good enough or sufficient*, and *least/dropped, or needs revising.*

Nurung (2008) conducted research in Kendari city. He found that the test reliability index is 0.826, there are 24 test items (60%) in good category and 16 test items (40%) are not in good cetegory so that the overall test quality is not quite good. Based on items responses theory using the BIGSTEPS program it is found that the test information function is 0.838 which means the test is reliable. There are 35 test items (87.5%) in good category and 5 terms (12.5%) is not good category to make the overall test quality falls into good category. The total number of good test items based on the three of analysis methods of analysis is (47.5%), while the bad test items are 21 (52.5%).

There is a lot of research that has been conducted by using ITEMAN program. From the related studies above, those studies mention that the researchers use ITEMAN as a tool to analyze multiple choice tests in elementary school, senior high school, and university as the population and sample of the research. This research will concern about reliability, level of difficulty, discriminating power, proportion of the answers and distractors. Because of that, the researcher analyze on those sides and investigating the population which has different knowledgeable students and multiple choice items, as the focus of this research. The researcher chooses the different population and sample that is in Junior high school at SMPN 2 Sumberejo, to know the differences of the result of the test.

2.2. Review of Related Literature

For the specific explanation about the analysis of final semester test using ITEMAN software program, the researcher explains some related literature about quality of a test, final semester test, multiple choice tests, guidelines for constructing multiple choice items, ITEMAN software program and assessment of multiple choice tests using ITEMAN software program, to know the quality of the final semester test such as validity, reliability, level of difficulty, discriminating power and distractors.

2.2.1. Quality of Test

One commonly used tool in assessment is a test. That is to assess the outcome of the learning process. To determine the quality of the test, it is necessary to analyze the test before the test is given to the participants of the test. According to Arikunto (2006:205), item analysis is a systematic procedure, which will provide information that is very specific to the test items arranged.There are two approaches that can be used to determine the quality of a test, namely qualitative and quantitative approaches (Osterlind, 1998:84). A qualitative approaches done by reviewing items and should be done before the test is tested. The thing which is emphasized is the assessment from the aspects of material, construction, and language. While the quantitative approach is a method of test item review based on empirical data obtained through participant responses. Item characteristics are a quantitative parameter.

The test should be qualified, because if the test is qualified, the result of the test will be good and reflect the students have been learned before. The quality of the test is important to separate between high, moderate and low level of the students. A good test can present students' achievement well. According to Athiyah (2012) a test can be said as a good test if it fullfils several requirements of a good test, both statistically and non statistically.

In ITEMAN software program, the measurement of validity is not covered explicitly. To know the validity of a test using ITEMAN, the value covers the level of difficulty, discriminating power, and proportion of the alternatives (Salirawati, 2011:28). Then, the conclusion from the three aspects gives a decision whether the test has good validity or not.

There are three types of validity used in this research: construct validity, content validity, and face validity. This research uses these types of validity due to that fact that in ITEMAN, the validity is not statistically computed. Consequently, construct validity, content validity, and face validity help the researcher determine the validity more accurately.

1) **Construct Validity**

The underlying theoretical construct in a test is concerned in this validity. The term "construct validity" refers to the overall construct or trait being measured (O'Neill, 2009:26). If a test is supposed to be testing the construct of speaking, it should indeed be testing speaking, rather than listening, reading, writing, vocabulary, and grammar. Therefore, the term construct validity has been used both for correspondence at the element level and at the relation level (Brinberg& McGrath, 1985:115).

a. *Traits of Reading*

Reading deals with how the readers receive the meaning through the written symbols and process them into their mind. Reading is one of the important skills which are needed by students from elementary school to university. Heaton (1975:105) states reading as recognizing words and word groups, associating sounds with responding graphic symbols. He defines reading comprehension as the questions which are set to test the students' ability to understand the gist of a text and to extract key information on specific points in the text. It indicates that comprehending the reading text involves connecting information from the written message to arrive at the meaning of the text.

Comprehension is very prominent in this case. Because of that, traits of comprehending texts which are evaluated indirectly put a heavier burden on the testing procedures which the tester decides to use and may have an effect on the score of the test taker (Shohamy, 1985:103).

To find the construct validity of the reading test, the final semester test was formulated by the concept of reading comprehension. According Davenport (2007: 61), common types of questions found in reading comprehension are included as follows:

1. Identifying main idea, main point, author purpose or an alternate title for the passage.
2. Recognizing the tone of the passage or identify the style.

3. Comprehending information directly stated in the passage (finding supporting detail).
4. Answering relational questions about the author’s opinion or idea, even if not stated directly.
5. Recognizing the structural methodology employed to develop the passage, for example sequence, vocabulary, and represent pronoun (reference).
6. Extending limited information given by the author to a logical conclusion using inference (inference meaning).

This research is focused on main idea, supporting detail, inference meaning, vocabulary, and reference.

b. *Traits of Grammar*

Grammar is one of the language components. In testing grammar, multiple choice test is one of the most common types. To test awareness of the grammatical features of the language using the objective test (multiple choice test), the test evaluates the ability to recognize or produce correct forms of language rather than the ability to use language to express meaning, attitude, emotion, etc (Heaton, 1975:34*).* It refers to pattern of form and arrangement by which the words are put together, because, according to De Capua (2008:1), grammar is a set of rules. One must also know how the words work together in English sentences, not only knowing English words and their meanings (Allen, 1983:2). Therefore, someone using language has to know the grammatical pattern of the language.

c. *Traits of Vocabulary*

If students cannot master vocabulary, they will fail to use the language both in oral or written form. Therefore, in order to be able to master the language, the students must learn vocabulary well. Not only a certain number of vocabularies, but they also know all vocabularies in order to master the language and use the words properly in vocabulary testing. Wallace (1986:1) states that vocabulary is the vital element of the language. As stated by Heaton (1975:51), vocabulary tests are designed that they test knowledge of words which, though frequently found in many English textbooks, are rarely used in ordinary speech. Subsequently, a careful selection, or sampling, of lexical items for inclusion in vocabulary test is the most crucial task.

2) **Content Validity**

Content validity represents the correlation between the test and exact materials, in terms of construction. As known that content validity is concerned with identifying the relationship between test tasks and specific learned content, construct validity attempts to make the connection between test tasks and theoretical constructs of language proficiency regardless of learned materials (Azwar, 2000:45). In the case of semester test, of course, there are no test specifications, and the teachers may simply need to check the teaching syllabus or the course textbook to see whether each item is appropriate for that examination.

3) **Face Validity**

Although this validity is considered as a weak measure, its importance cannot be underestimated. Face validity is very important for holistic scores. Holistic tests that measure writing look at actual pieces of writing to do so (Lynne, 2004:35). According to O’Neill (2009:26), face validity is a test looked like it would measure the desired ability or trait. So, if the test lacks face validity, it may not work as it should, and may have to be redesigned.

a. Reliability (Chronbach Alpha)

If the results of a test are replicated consistently, they are reliable. In psychometrics, reliability is a technical measure of consistency (Lynne, 2004:31).Test reliability is very important for a test user because it is necessary for good validity. In short, a test can be highly reliable without necessary being valid for any purpose of interest. Test reliability refers to the reproduced ability of test results. A test with high reliability is one that will reproduce very much the same relative important of test score for a group of students under different condition or situations.

There are three indexes that can be followed to determine whether the reliability of a test is low, average and high as follows:

Table 2.1 Criteria of Reliability (Alpha)

Criteria	**Index**	**Clasification**
Reliability (Alpha)	0,000 - 0,400	Low
	0,401 - 0,700	Average
	0,701 - 1,000	High

Source: Suparman (2011)

2.2.2. Final Semester Test

Final semester test is an activity that is carried out by educators to measure students' achievement on competencies at the end of the semester. The test comprises all indicators that represent all of the standard competence in the semester (Permendiknas No. 20, 2007 on the Standard Assessment).Based on the article, it asserts that the final semester test given by educators is under the coordination of the educational unit. Because of that, the educators or teachers have to conduct an assessment of their students under the coordination of the school as an educational unit. The provisions indicate that the teachers have an important role to determine the progress of the students through final semester test. This is relevant to the evaluation of the characteristics of education where the most ideal in evaluating education is teacher as an educator.

2.2.3.Multiple Choice Tests

Multiple choice test items are designed to elicit specific responses from the students. Over the last decade, large student numbers, reduced resources and increasing use of new technologies have led to the increased use of multiple choice questions as a method of assessment in higher education courses (Nicol, 2007:53). According to Wiggins & McTighe (2005:338), multiple choice tests are indirected measures of performance. For those students who have not achieved the objectives, the distracters appear as plausible solutions to them. On the contrary, only the answer should appear plausible to these students and the distracters must emerge as implausible solutions for those students who have achieved the objectives.

The alternatives may be complete sentences, sentence fragments, or even single words. In fact, the multiple choice items can assume a variety of types, including absolutely correct, best answer, and those with complex alternatives (Osterlind, 1998:20). It is supported by Hughes (2005:75) who states the most obvious advantage of multiple-choice is that scoring can be perfectly reliable. In line with Hughes, Valette (1967:6) states that scoring in multiple choice techniques is rapid and economical and it is designed to elicit specific responses from the students.

Multiple choice can be used any condition and situation, in any level or degree of education. Multiple choice items take many forms but their basic structure is that has stems or the questions itself, and a number of options-one which is correct, the others being distractors (Hughes, 2005:75).

Multiple choice test items are designed to elicit specific responses from the students. Since there is only one right answer, the scorer can very rapidly mark an item as correct or incorrect. More important, when a group of scorers is reading the same test paper, each of them arrives at the same score. although the scorer reliability of multiple choice test is almost perfect, the validity of each test or each section of the test must be determined separately. Just because a test is objective, it is not automatically a good test. Before using any standarized objective test, the teacher should carefully go over the specifications to determine whether they correspond to his or her own reasons for giving the test.

2.2.4. Guidelines for Constructing Multiple Choice Items

According to Gronlund (1968:92) the multiple choice item consists of a stem, which presents a problem situation, and several alternatives (options or choices), which provide possible solutions to the problem. The stem may be a question or an incomplete statement. The alternatives include the correct answer and several plausible wrong answers called distracters. The function of the latter is to distract those students who are uncertain of the answer. There are several rules for writing multiple choice items:

1. design each item to measure an important learning outcome,
2. present single clearly formulated problem in the stem of the item,
3. state the stem of the item in simple, clear language,
4. put as much of the wording as possible in the stem of item,
5. state the stem of the item in positive form, wherever possible,
6. emphasize negative wording whenever it is used in the stem of an item,

7. make certain that the intended answer is correct or clearly best,
8. make all alternatives grammatically consistent with the stem of the item and parallel in form,
9. avoid verbal clues that might enable students to select the correct answer or to eliminate an incorrect alternative,
10. make the distracters plausible and attractive to the uninformed,
11. vary the relative length of the correct answer to eliminate length as a clue,
12. avoid using the alternative “all of the above,” and use “none of the above” with extreme caution,
13. vary the position of the correct answer in a random manner,
14. control the difficulty of the item either by varying the problem in the stem or by changing the alternatives,
15. make certain each item is independent of the other items in the test,
16. use an efficient item format,
17. follow the normal rules of grammar, and
18. break (or bend) any of these rules if it will improve the effectiveness of the item.

There are some steps to guideline the test items. All of the steps are important and good for test maker or the teachers to make the good and qualified test items. But, the most important steps are the design each item to measure an important learning outcome, state the stem of the item in simple and clear language, make certain that the intended answer is correct or clearly best, make all alternatives grammatically consistent with the stem of the item and parallel in form, control the difficulty of the item either by varying the problem in the stem or by changing the alternatives and follow the normal rules of grammar.

2.2.5. ITEMAN Software Program

ITEMAN is an accurate software program with the beginning stamping back to the 1960s(Nelson,2012).For quite a few years, it was designed to be utilized for traditional item and test analysis. As a complete and reliable workhorse, it has had decades to solidify notoriety. The ITEMAN software program is publicized as a Classical Item Analysis program. Not only to estimate and note test scores, but also can examine multiple choice questions. The model of the program is 3.50, at hand on the internet at www.assess.com. There are four statistical measures offered in the program (ASC, 1989-2006:13):Proportion Correct, Discrimination Index, Biserial and Point Biserial Correlation Coefficients.

ITEMAN is very important for teachers of English an all levels (Junior, Senior and college) who have to be responsible for administering tests, such as mid semester or final examination, so that they can be sure about the quality of the test item that they will use. According to Suparman (2011 : 86), the data, which have been input or keyed into the computer to be analyzable by ITEMAN, should be formatted in ASCII (text only) files. One of the advantages of the ITEMAN, according to Suparman (2011 : 86), is that a single analysis can accommodate up to 750 items, while the number of the examinees is almost unlimited.

Here are brief descriptions of the research's commonly used terms, to allow for better understanding when they appear in the remainder of the paper. All these formulas are not used in practice because ITEMAN analyzes them automatically.

1. Proportion Correct (Level of Difficulty – p)

Level of difficulty is simply the precentage of students taking the test who answered the item correctly. The larger precentage getting an item right, the easier item. Probably the most popular item-difficulty index for dichotomously scored test or multipoint items is the p-value (Osterlind, 1998:266). It is simply the proportion (or percentage) of students taking the test who answered the item correctly (Haladyna, 2004:207).

Here are indexes that can be followed to determine whether a test, as follows:

Table 2.2 Criteria to classify the quality of test items level of difficulty

Index	Decision
0.000 – 0.099	Very difficult/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/needs dropping or total revising

Source: Suparman (2011)

2. Point Biserial (Discriminating Power – D)

Discriminating power refers to the capacity of a test to discriminate between the clever and the stupid students. There are two indicators of the item discrimination effectiveness, which a point biserial correlation and biserial correlation coefficient (Matlock-Hetzel, 1997). The size of the discrimination index is informative about the relation of the item to the total domain of knowledge or ability, as represented by the total test score (Haladyna, 2004:211). The advantage of using discrimination coefficient over the discrmination index is that every person taking the test is used to compute the discrimination coefficient. The point biserial correlation is applied to determine whether the right people are getting the items right, and how much predictive power the item has and how it would contribute to prediction.

There are four indexes that can be followed to determine whether a test item is very low, low, average and high as follows:

Table 2.3 Criteria of Discriminating Power (D)

Index	**Clasification**	**Decision**
D ≤ 0,199	Very low	Rejected/total revising
0,200 - 0,299	Low	Revised
0,300 - 0,399	Average	Accepted
D ≥ 0,400	High	Accepted

Source: Suparman (2011)

3. Prop Endorsing (Propotion of The Answer)

According to Supranata (2006), prop endorsing (propotion of the answer) is considered functional if at least chosen by 5% of the examines. Propotion of the answer is one of the important things to determine the quality of multiple choice tets. The test item is good if the propotion of the answer has minimum index 0.401, it is average or sufficient. If the propotion of the answer less than 0.401 it means need revising or should be dropped.

There are indexes that can be used to decide the prop endorsing (propotion of the answer) test item, as follows:

Table2.4 Prop Endorsing (Propotion of the Answer)

Index	**Decision**
0.000 – 0.010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very good

Source: Suparman (2011)

2.2.6. Advantages and Disadvantages of ITEMAN

ITEMAN is one of the new program in assessing students ability. As a new program, ITEMAN has some advantages and disadvantages. There are some advantages and disadvantages of ITEMAN software program:

ITEMAN has some advantages in assessing students ability. There are some advantages ITEMAN software program as follows:

1. ITEMAN is simply to apply

   ITEMAN is easy to use and very simple, we just need an electricity, computer and ITEMAN software program we have been use ITEMAN. ITEMAN can be used everywhere, anywhere and for everyone.

2. ITEMAN easy to understand

   Everyone can use ITEMAN, because steps to use ITEMAN is very easy and simple. We just need to follow the steps and we automatically can use ITEMAN.

3. ITEMAN can minimize the time

   By using ITEMAN the teacher can analyzing up to 750 data. After we input the data to the computer, we just need one click to see the result of our anlysis.

4. ITEMAN make the teacher easier to assess the students

   ITEMAN can be used to determine the validity, reliability, level difficulty, point biserial, discriminating power and key answer. By using ITEMAN teacher will be easier to assess the students ability.

ITEMAN has some advantages, but beside that ITEMAN also has some disadvantages. There are some disadvantages of using ITEMAN as follows:

1. ITEMAN can be used if in one school has electricity connecttion.

   If in the school there is no electricity connection ITEMAN can not be used.

2. ITEMAN just can be used if in the school has been used computer. Because ITEMAN is a software program, so the school should have computer to access the program.

3. ITEMAN can be used if we can operate the computer.

Because ITEMAN is a software program in computer, we need computer to access the ITEMAN software program.

2.2.7. Assessing Multiple Choice Tests Using ITEMAN Program

When the test analyzed by ITEMAN is composed of multiple scales, the items are assigned to the scales using the inclusion codes. This means that statistics analysis about the test is provided in the output data of ITEMAN. Particularly, the exemption of file capability in ITEMAN gives an opportunity to the examinees to re-analyze data of the multiple choice tests if students find that they want to take into account of more than one option/alternative as the correct (keyed) alternative. Some possible circumstances for giving credit to more than one alternative include poorly phrased questions, conflicting source information, or an indication of additional problems from a previous analysis. No single response is considered correct and the item has no influence on the total score (ASC, 1989-2006:3).According to Surapranata (2006), an alternative is considered functional if at least chosen by 5 % of the examinees.

ITEMAN analyzes scales containing either dichotomously score or multipoint items. The program can work only with multiple choice items. It is relatively easy to analyze test items using the ITEMAN program.

CHAPTER 3

RESEARCH METHOD

In this chapter, the method of the research is discussed. The parts of methodology such as: setting of the research, research design, population and sample, data collecting techniques, research procedure, data analysis, and hypothesis testing are explained further.

3.1. Setting of the Research

The researcher chose SMP NEGERI 2 Sumberejo Tanggamus as the research place because this is one of the developing schools in Tanggamus, Lampung province that can be reached easily by the researcher. The researcher was tested the final semester test items at first semester students class VIII. And for the time of research, researcher prepared for the proposal, determined the object of the research, determinedthe subject, approached the school and the teachers, seeked for permission from the headmaster and teachers of English to carry out the research in that school for one specific time.

3.2. Research Design

This research was descriptive quantitative. This researchwas intended to search for the information about the quality of final semester test in class VIII SMPN 2 Sumberejo in 2016/2017 academic year. This research used descriptive quantitative approach where explanation using descriptive methods and the result processed by using ITEMAN software program. The quantitative approach in this research was by using ITEMAN software program. After the data was collected, the data had been processed by using ITEMAN. The result of the data had been explained descriptively to know the quality of the test items, whether it was good or not.

3.3. Population and Sample

The population of this research was all of the test items in SMP N 2 Sumberejo Tanggamus at the second grade. The researcher chose one test item that was English test item because the researcher concerned with English education. The subject of this research is ITEMAN and the object is final semester test at the second grade of SMPN 2 Sumberejo. There were five classes in VIII grae of SMPN 2 Sumberejo. The researcher chose class VIII 3 because this class is the middle class in grade VIII. In this class, there were smart, average and low level students. So, this class was very availble as research object.

3.4. Data Collecting Technique

The data were collected from the final semester test. There were fifty questions. The researchertook the students' answers and the test from the school. Then, ITEMAN software program gets its turn to analyze the test.

3.5. Research Procedures

The researcher checkedthe quality of the final semester test after the students' answers and question sheets have been obtained. The instrumentwas the final semester test; each item had five options A., B., C., D., and E. Then, the researcher analyzed the test.

There were several procedures to make the research run well, as follows:

1. Determining the problems

   The problems were formulated to be a foundation of this research.

2. Determining and selecting the population

   The population of this research wasall of the final semester test items at the second grade of SMP N 2 Sumberejo Tanggamus.

3. Determining the data

   The researcher took one class. The sample of this research was class VIII3.

4. Determining the test

   The test was from the final examination. There were fifty test items.

5. Carry out the test

The studentshave been given the test and answer the questions. The students have been given 60 minutes to answer the multiple choice test items.

6. Collecting the test

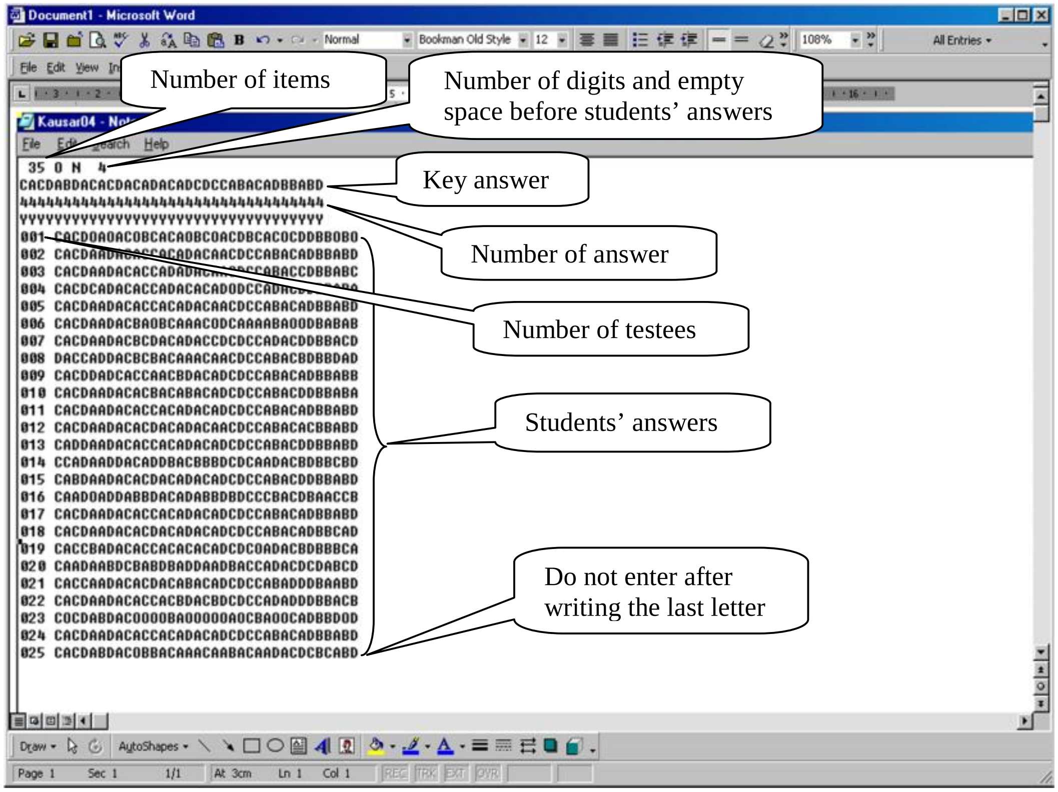
   After students answer the questions, the researcher collected the answer sheets of the students.

7. Analyzing the data quantitatively

   This research touchedthe final semester test by counting on ITEMAN software program.

8. Analyzing the data descriptively

   Final semester test had been identified by using descriptive approach to find out the reliability, level of difficulty, discriminating power, and quality of the distractors.



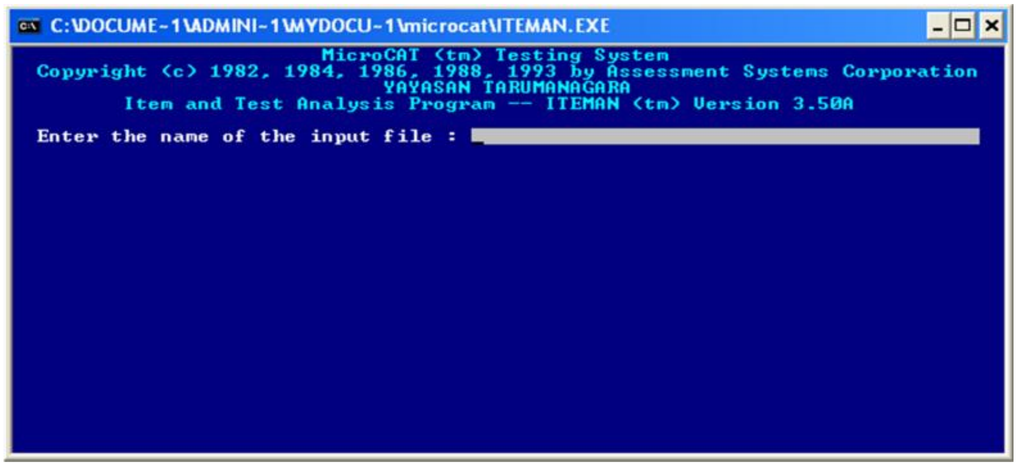
```
 35 O N  4
CACDABDACACDACADACADCDCCABACADBBABD
44444444444444444444444444444444444
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
001 CACDOAOACOBCACAOBCOACDBCACOCDDBBOBO
002 CACDAADACACCACADACAACDCCABACADBBABD
003 CACDAADACACCADADACAACDCCABACCDBBABC
004 CACDCADACACCADACACADODCCADACDDBBABA
005 CACDAADACACCACADACAACDCCABACADBBABD
006 CACDAADACBAOBCAAACODCAAAABAOODBABAB
007 CACDAADACBCDACADACCDCDCCADACDDBBACD
008 DACCADDACBCBACAAACAACDCCABACBDBBDAD
009 CACDDADCACCAACBDACADCDCCABACADBBABB
010 CACDAADACACBACABACADCDCCABACDDBBABA
011 CACDAADACACCACADACADCDCCABACADBBABD
012 CACDAADACACDACADACAACDCCABACACBBABD
013 CADDAADACACCACADACADCDCCABACDDBBABD
014 CCADAADDACADDBACBBBDCDCAADACBDBBCBD
015 CABDAADACACDACADACADCDCCABACDDBBABD
016 CAADOADDABBDACADABBDBDCCCBACDBAACCB
017 CACDAADACACCACADACADCDCCABACADBBABD
018 CACDAADACACDACADACADCDCCABACADBBCAD
019 CACCBADACACCACACACADCDCOADACBDBBBCA
020 CAADAABDCBABDBADDAADBACCADACDCDABCD
021 CACCAADACACDACABACADCDCCABADDDBAABD
022 CACDAADACACCACBDACBDCDCCADADDDBBACB
023 COCDABDACOOOOBAOOOOOAOCBAOOCADBBDOD
024 CACDAADACACCACADACADCDCCABACADBBABD
025 CACDABDACOBBACAAACAABACAADACDCBCABD
```

Figure 3.1. An example of Data file using notepad on Windows

Source: Suparman (2011)

After all the data have been put in Notepad and saved, the data were analyzed using ITEMAN program. The following were the steps of utilizing the program (Suparman, 2011):

1. Open ITEMAN Program, by clicking *Start*
2. Select *program*/click ITEMAN and the program shows this appearance.



```
                        MicroCAT (tm) Testing System
 Copyright (c) 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
                             YAYASAN TARUMANAGARA
          Item and Test Analysis Program -- ITEMAN (tm) Version 3.50A

 Enter the name of the input file :
```

3. Type the name of your data file (input) on *Enterthe name of the input file*. For example, D:\MIDTEST.txt, then *Enter*
4. Enter the name of the output file on *Enter the name of the output file.* For example, D:\MIDTEST.output, then click *Enter*
5. A question appears*Do you want the scores written to a file? (Y/N)*, then type *Y* and click *Enter*.
6. Enter the name of your score file on *Enter the name of the score file*: For example, D:\MIDTEST.scr, then click *Enter*, *Finish*.

Then, there are some steps to open the results of item analysis on MS WORDS program (Suparman, 2011):

1. Click *Start*
2. Select *Program*/click *Microsoft Word*
3. Click *File*/click *Open*, then look for the results on, for example, *Drive D* (depends on which one you choose)
4. The result appears in output data.

The previous statistics are provided by ITEMAN for each scale (subtest) analyzed (ASC, 1989-2006:16-18):

1. **N of Items.** The number of items in the scale that are included in the analysis.
2. **N of Examinees.** The number of examinees that are included in the analysis for the scale.
3. **Mean.** The average number of items on each scale that were answered correctly.
4. **Std.Deviation.** Deviation standard from score distribution of the examinees.
5. **Minimum.** The lowest score on each scale for any examinee.
6. **Maximum.** The highest score on each scale for any examinee.
7. **Alpha.**Cooefisien realibility alpha for the test which is homogeneus of the test/scale.
8. **Mean Biserial.** Mean of the differenciate score index by counting from the mean of biserial correlation.

The statistic in output data of ITEMAN was used when the rater identifies a mastery criterion group within the group of students being tested. The upper scoring group was usually the group that passes the test; whereas the lowerscoring group wasusually the group that failed the test. To find out the proportion correct, the statistic was calculated by taking the number of master students answering the item correctly, subtracting the number of non-master students answering the item correctly and then dividing by the total number of students (Crocker &Algina, 1986). It is calculated by the following formula (Backhoff, Larrazolo, & Rosas, 2000):

`

$$pi = \frac{Ai}{Ni}$$

where:

pi = Difficulty index of item *i*

Ai = Number of correct answers to item *i*

Ni = Number of correct answers plus number of incorrect answers to item *i*

The program can process up to a 750-item test with unlimited number of students(ASC, 1989-2006).The user can also manually created a data file using the edit menu in ITEMAN, which was similar to Windows Notepad program. ITEMAN’s controls were few in number and very simple to use. The program offeed five pull down menus and five buttons. The user first selects the configure menu or button to identify the file and select the options desire for analysis. The user then selected the analyze menu or button. The user could viewed or print the output file by clicking on the view button or print button. These buttons appear after the analysis was complete.

CHAPTER 5

CONCLUSIONS AND SUGGESTIONS

This chapter deals with the conclusions and the suggestions based on the results and the discussions of this research.

5.1. Conclusions

The findings of the research specify that not all items in the final semester test have high reliability, average level of difficulty, high discriminating power, and good prop endorsing. Some of the test items have low reliability, low level of difficulty, level of difficulty, low discriminating power and least prop endorsing. It means that not all of the test items are proper enough to test students ability, there are some mistakes and some revisions needed to make the test items properly.

In the output data of the ITEMAN, the result shows that the reliability coefficient of alpha is 0.592. Based on the criteria of the reliability of the test items, it is categorized as average/sufficient, that is, the test items whose alpha ranges from 0.401 – 0.700. It means that the test items in general if they are tested frequently under the same condition, they might result in similar outcome.

The test items are good if they are in average level. So, if the test is in the average level of difficulty, the test is good for the students. Related to the result of the level of difficulty in the output data of ITEMAN, some of the items fulfill the quality of a good item, but some do not.

Regarding with the item analysis using ITEMAN, it was found that the level of difficulty can be classified into five categories, that is, *very difficult/needs total revising, difficult/needs revising, average/good, easy/needs revising and very easy/needs dropping or total revising.* For the criteria *very difficult or needs total revising,* the items have the level of difficulty ranging from 0.000-0.099. This class consists of 5 items (10%). There are five items that are very difficult, that is, 6, 7, 11, 25, 48. These items need to be total revised. With reference to the criteria of the items which have the level of difficulty ranging from 0.100 – 0.299, the items are categorized as *difficult/needs revising*. This class consists of 16 items (32%). There are sixteen items that are *difficult*, that is, 4, 14, 17, 21, 22, 26, 27, 28, 31, 32, 33, 36, 37, 38, 44, 46, therefore, they need revising. The criteria of the items which have the level of difficulty ranging from 0.300-0700 are categorized as *good or directly usable*. This class consists of 22 items (44%). There are twenty two items that are *good*, that is 1, 2, 5, 9, 12, 15, 16, 18, 20, 24, 29, 30, 34, 35, 40, 41, 42, 43, 45, 47, 49, 50. These items also need to be totals revised. With reference to the criteria of the items which have the level of difficulty ranging from 0.701 – 0.900, the items are categorized as *easy/needs revising*. This class consists of 5 items (10%). There are five items that are *easy*, that is, 8, 13, 19, 23, 29, therefore, they need revising. These items are recommended to be directly used without any prior revision. As to the category *very easy or needs total revising*, the items have the level of difficulty ranging from 0.901-0.1000. This class consists of 2 items (4%). There are two items that are very easy, that is, 3, 10.

Related to the item analysis using ITEMAN, it was found that the test items whose discriminating power ≥ 0.400 is classified as *high*. There are 13 items (26%) that are *high*, that is, 6, 9, 16, 17, 18, 19, 30, 35, 39, 40, 41, 42, 43. These test items are recommended to be used as they can discriminate between the more knowledgeable from the less knowledgeable students. The criteria *average/without revising* is the items whose discriminating power ranges from 0.300-0.399. There are 2 items (4%) that do not need revising, that is, 2, 44. Concerning with the criteria *low/needs revising*, it points out that the items whose discriminating power ranges from 0.200-0.299. It was found that there are 12 items (24%) which involve in low discriminating power or need to be revised, that is, 5, 10, 11, 20, 20, 22, 23, 29, 31, 32, 34,47. The test items whose discriminating power range from 0.000-0.199 are categorized as *very low/needs dropping*. There are 23 items (46%) that are *total revising,* that is, 1, 3, 4, 6, 7, 13, 14, 15, 21, 24, 25, 26, 27, 28, 33, 36, 37, 38, 45, 46, 48, 49, 50.

Based on the results of the data analysis using ITEMAN, it was found that the alternative or prop endorsing of the 50 items consisting of A, B, C and D with the total of the alternatives is 200, can be classified into three categories, that is, *very good, good enough or sufficient,* and *least/dropped, or needs revising*. With respect to the criteria *very good*, the alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.051-1.000. This class consists of 26 options (15%). These alternatives are recommended to be used without any prior revision. The alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.011-0.050 is categorized as *good enough or sufficient*. This class consists of 43 options (24.5%). These alternatives are recommended to be directly used, because they are chosen by at least 5% of the testees. Related to the criteria *least/dropped, or needs revising*, it is the alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.00-0.010. This class consists of 46 options (60.5%). These items should be revised before being tested.

5.2. Suggestions

In line with the conclusions above, some suggestions are proposed as follows:

1. Suggestions to the teachers
   a. The teacher should be good at the assessment from the aspects of material, construction, and language in order to improve the quality of the test.
   b. The teachers should be familiar with ITEMAN software program in order that they can assess the students' ability faster.
   c. The teachers should be trained to use ITEMAN software program in order to improve the quality of the test.
   d. The teachers should be familiar with all the terms related to the quality of the test items, such as, validity, reliability, prop. Correct (level of difficulty), point biserial (discriminating power), prop. Endorsing (options), distracters, key answers, alpha, and standard deviation.

2. Suggestions to other researchers
    a. It is suggested that the role of ITEMAN in determining the quality of multiple choice items is investigated further. It is also interesting to collect a larger or smaller data base for investigating whether there are more tendencies in determining the quality of items.
    b. Other researchers should replicate the current study in analyzing the quality of other test items, such as, Final School Test (UAS), and National Examination (UN).

REFERENCES

Al-Ghani, Bagus. (2014). *Analyzing the Quality of the Final Semester Test Using Iteman Software Program at the Second Year of SMA Negeri 1 Purbolinggoin 2013/2014 Academic Year*. (Script). University of Lampung.

Arifin, Zainal. (2011). *Evaluasi Pembelajaran*. Bandung: PT Remaja Rosdakarya

Assesssment Systems Corporation. (1989–2006). *User's Manual for the ITEMAN^{TM} Conventional System Analysis Program*. 2233 University Avenue, Suite 200.

Arikunto Suharsimi. (2006). *Prosedur Penelitian Suatu Praktik*. Jakarta: Rineka Cipta.

Bakhoff, E., Larrazolo, N., & Rosas, M. (2000). The level of difficulty and disrimination powerof the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electronica de Investigation Educativa, 2.*

Creswell, John W. (2009). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Los Angeles: Sage.

Daryanto. (2008). *Evaluasi Pendidikan*. Jakarta: PT Rineka Cipta

Fitriana, Novaria. (2013). *Analisis Kualitas Butir Soal Ulangan AKHIR Semester Gasal Mata Pelajaran IPA Kelas V Mi Sultan Agung Tahun Pelajaran 2012/2013*. (Skripsi). Universitas Islam Negeri Sunan Klijaga, Yogyakarta.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items 3rd ed*. New Jersey: Lawrence Elbaum Asociate.

Hughes, A. (2005). *Testing for Language Teachers*. 2nd Ed. London: Cmbridge University Press.

Lynne, Patricia. (2004). *Coming to Terms: Theorizing Writing Assessment in Composition Studies*. Logan: Utah State University Press.

Matlock-Hetzel, S. (1997). *Basic concepts in Item and Test Analysis*. Texas.: A&M University.

Muh, Nurung (2008). *The Quality of the Final Examination Test of IPA SD of National Standard School in the Academic Year of 2007/2008 in Kendari City of South East Sulawesi*. Thesis. Yogyakrta: Graduate School, State University of Yogyakarta.

Nelson, Lary. (2012). *ITEMAN 2 and Lertap 5*. Curtin University of Technology.

Nicol, David. (2007).E-assessment by Design: Using Multiple-Choice Tests to Good Effect. *Journal of Further and Higher Education, Vol. 31, No. 1, February 2007, 53-64.*

O'Neill, Peggy. (2009). *A guide to College Writing Assessment*. Logan: Utah State University Press.

Osterlind, Steven J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats*. University of Missouri-olumbia.

Ratnaningsih, Dewi J. (2009). *Analisis Butir Soal Pilihan Ganda Ujian Akhir Semester Mahasiswa Di Universitas Terbuka Dengan Pendekatan Teori Tes Klasik*. (Skripsi). Universitas Terbuka, Tangerang.

Sanjaya, Wina. (2008). *Perencanaan dan Desain Sistem Pembelajaran*. Jakarta: Kencana Prenada media group.

Surapranata, Sumarna. (2006). *Analisis Validitas, Reaibilitas dan Interpretasi Hasil Tes Implementasi Kurikulum 2004*. Bandung: PT Remaja Rosda Karya.

Suparman, U. (2011). The Implementation of Iteman to Improve the Quality of English Test Items As a Foreign Language: *An Assessment Analysis AKSARA-Jurnal Bahasa, Seni dan Pengajarannya,* Vol-XII, No. 1, pp 86-96.

Wiggins, G. & McTighe, J.(2005). *Understanding by Design (Expanded 2nd Ed. USA)* Alexandria, Va: Association for Supervision and Curriculum Development.