EVALUATING ENGLISH FINAL SEMESTER TEST MADE BY THE TEACHER BY USING ITEMAN SOFTWARE PROGRAM AT SMAN 5 BANDAR LAMPUNG

(A Script)

By
Saghina Meividia Anas

FACULTY OF TEACHER TRAINING AND EDUCATION
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2019

ABSTRACT

EVALUATING ENGLISH FINAL-SEMESTER TEST ITEMS MADE BY THE TEACHER BY USING ITEMAN SOFTWARE PROGRAM AT SMA N 5 BANDAR LAMPUNG

BY

SAGHINA MEIVIDIA ANAS

The objectives of this research were to determine the quality of final test made by the teacher based on the criteria of a good test: (1) validity, (2) reliability, (3) level of difficulty, (4) discrimination power, (5) the quality of the alternatives at SMAN 5 Bandar Lampung at X SOS 4 class. This research applied qualitative descriptive approach which report the criteria of the test. The data were taken from the students' answer sheet and the multiple choice question which was used for final semester test by using iteman software program.

The result of the research showed that (1) The validity of the test was not proper enough to used because it did not fulfil the requirements, (2) The reliability was sufficient, based on scale statistics, that is 0.492, (3) The level of difficulty of the test items consisted of 16 (35.56%) test items considered good, 16 (35.56%) very difficult, 11 (24.44) items difficult, 1 (2.22) easy and 1 items very easy, (4) The discrimination power of the test items consisted of 10 (22.22%) categorized as high, 5 (11.11%) average, 2 (4.45%) as low/need revising, and 28 (62.22%) grouped into very low or need dropping, (5) The qualities of the alternatives consisted of 95 alternatives (42.22%) should be dropped, 113 alternatives as good distractors, and 17 as very good distractors.

In addition, based on the output data of iteman software program, there were some items that should be revised by changing the key answer and several numbers that can be used without any revision because they provided the proper anwer key for the test. It can be concluded that the quality of the test items was moderate.

Keywords : *evaluating, validity, reliability, level of difficulty, discrimination power*

EVALUATING ENGLISH FINAL SEMESTER TEST MADE BY THE TEACHER BY USING ITEMAN SOFTWARE PROGRAM AT SMAN 5 BANDAR LAMPUNG

By
Saghina Meividia Anas

A Script

Submitted in a Partial Fulfillment of
The Requirements for S-1 Degree

in

The Language and Arts Department of
Teacher Training and Education Faculty

FACULTY OF TEACHER TRAINING AND EDUCATION
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2019

Research Title	:	EVALUATING ENGLISH FINAL TEST MADE BY THE TEACHER BY USING ITEMAN SOFTWARE PROGRAM AT SMAN 5 BANDAR LAMPUNG
Student's Name	:	**Saghina Meividia Anas**
Student's Number	:	**1513042011**
Department	:	**Language and Arts Education**
Study Program	:	**Teacher Training and Education**

APPROVED BY
Advisory Committee

Advisor	Co-Advisor
Prof. Dr. Cucu Sutarsyah, M.A	**Dr. Muhammad Sukirlan, S.Pd., M.A.**
NIP 19570406 198603 1 002	NIP 19641212 199003 1 003

The Chairperson of
The Department of Language and Arts Education

Dr. Nurlaksana Eko R., M.Pd.
NIP 19640106 198803 1 001

ADMITTED BY

1. Examination Committee

   Chairperson : **Prof. Dr. Cucu Sutarsyah, M.A.** ……………

   Examiner : **Drs. Ujang Suparman, M.A., Ph.D.** ……………

   Secretary : **Dr. Muhammad Sukirlan, S.Pd., M.A.** ……………

2. The Dean of Teacher Training and Education Faculty

Prof. Dr. Patuan Raja, M.Pd.
NIP 19620804 198905 1 001

Graduated on: **October 14th, 2019**

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini, saya:

Nama	: Saghina Meividia Anas
NPM	: 1513042011
Program Studi	: Pendidikan Bahasa Inggris
Jurusan	: Pendidikan Bahasa dan Seni
Fakultas	: Keguruan dan ilmu pendidikan
Judul Skripsi	: Evaluating English Final Semester Test Made by the Teacher by Using Iteman Software Program at SMAN 5 Bandar Lampung

Menyatakan bahwa skripsi ini adalah hasil karya sendiri. Sepanjang pengetahuan saya, karya ini tidak berisi materi yang ditulis oleh orang lain, kecuali bagian-bagian tertentu yang saya ambil sebagai acuan. Apabila ternyata terbukti bahwa pernyataan ini tidak benar, sepenuhnya menjadi tanggung jawab saya.

Bandar Lampung, Oktober 2019
Penulis,

METERAI TEMPEL
TGL. 20
5E85CAHF038804695
6000
ENAM RIBU RUPIAH

Saghina Meividia Anas

CURRICULUM VITAE

The writer's name is Saghina Meividia Anas. She was born in Bandar Lampung, on May 11th 1997. She is the second daughter from H. Saprul Huda, S.E., and Hj. Dra. Darnawati.

She began her formal educational institution for the first time at Dewi Sartika Kindergarten School in 2002. Then she continued her study at SDN 1 Sukarame in 2003 and graduated in 2009. In the same year, she registered at SMPN 5 Bandar Lampung. After graduating from Junior High School in 2012, she went to SMAN 5 Bandar Lampung and graduated in 2015.

In the following year, she continued her study at University of Lampung in 2015. She was registered as newly accepted student at the English Study Program, Language and Art Education Department through SNMPTN in 2015. In 2018, the researcher did teaching practice program (PPL) at SMPN 3 Raman Utara from July to August 2018. She did her research in SMAN 5 Bandar Lampung on May 2019. During her academic years, she was actively on an organization that was English Society (Eso).

DEDICATIONS

All praises and gratitude are only to Allah SWT, for all the tremendous blessings to me

This script is fully dedicated to:

My beloved parents

My beloved siblings

My special inspiration

My fabulous yet lovely best friends

My beloved lecturers at the English Department

My beloved comrades of English Department batch 2015

My beloved almamater, University of Lampung

MOTTO

Do good. And good will come to you.

-Anonymous-

The most simple things bring happiness.

-Saghina Meividia Anas-

ACKNOWLEDGEMENTS

Praise is only for Allah SWT, the Almighty God, for the mercy and blessing that enables the writer to finish the script. This script, entitled “Evaluating English Final Semester Test Items by Using Iteman Software Program at SMAN 5 Bandar Lampung” is presented to the Language and Arts Education Department at the Teacher Training and Education Faculty, University of Lampung as partial fulfillment of the requirements for S-1 degree in English Department.

It is important to be known that the script would never have come into existence without any supports, encouragements, and assistances by several generous people. The writer would like to take this opportunity to address her sincere gratitude and deep respect to:

1. Prof. Dr. Cucu Sutarsyah, M.A., as the first advisor, for his invaluable guidance, ideas, suggestions, and encouragements for the writer during the script writing process.
2. Dr. Muhammad Sukirlan, M.A., as the second advisor, who had contributed and given his evaluation, comments, suggestions during the completion of this script.
3. Drs. Ujang Suparman, M.A., Ph.D., as the examiner, for his encouragement and contribution during seminar until this script is finished.
4. Dr. Ari Nurweni, M.A., as the Head of English Department Study Program.
5. All lecturers of English Department who have given great contribution in broadening and deepening my knowledge during my study and to all staff members of English Department, Bu Sures, Mba Nur, and Mas Dwi who have helped me to organize my seminar.
6. Hendra Putra, S.Pd., M.Pd., as the headmaster of SMAN 5 Bandar Lampung and , mam Suwesi Erfina, S.Pd., M.Pd., as the English teacher of SMAN 5 Bandar Lampung who have given me the help and chance to conduct my research.
7. All students of SMAN 5 Bandar Lampung, especially classes X Sos 4 for their nice cooperation during this research.

8. My beloved parents, H. Saprul Huda, S.E., and Hj. Dra. Darnawati. Thank you for your endless love, for knowing my inside and outside, for teaching me to be serious in doing everything, for reminding me about my health, for supporting me to finish my script soon, and the most important, for always praying for me to be a good daughter.
9. My beloved brothers and sister, A. Yudha Prawira, S.H., Atha Afifah., and M. Vino Alfarabi, who always help, support, cheer, and remind the writer of how precious her struggle was.
10. My fabulous ISDAM, Tika, Lutfi, Melvy, Frilly, Hanny, Helda, and Shiane, who are never tired of supporting the writer to finish her study, always give trust, laughter, love and bullying.
11. My long life best friends, Rena and bella, who are absolutely encouraging the writer by sharing laughter and love.
12. Her big gratitude toED'15 squads that the writer cannot mention here.
13. My KKN Ratna Daya Family, Yanak cewe gila, Nola bundadari, Yanik wanita penghibur, Kia si manis manja, Kekem si paling rajin, Shinta si cempreng, Ghita nyonya persit, Azmi kordes cengeng and Ghalang sleeping beauty. thank you for your encouragement, cares, and jokes.
14. My mood booster, frienemy, lovely and best partner in life, Tegar Prasetio S.P (soon to be), thankyou for always inspiring, cheering, supporting and loving me in all conditions through these time.
15. My seniors and juniors in English Department, thank you for your greatest motivation, help and kindness.
16. My ESo team, especially Creativity and Financial Department, thank you for giving me such a great atmosphere during my college life.
17. Anyone who cannot be mentioned directly here who has contributed in complementing this script.

The researcher hopes that this research would be a positive contribution to the education development, the readers, and the other researchers.

Bandar Lampung, October 2019

The Researcher

Saghina Meividia Anas

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

Appendix

I. INTRODUCTION

This chapter deals with the background of the problems, identification of the problems, limitation of the problems, formulation of the research questions, objectives, the uses of the research, and the dfinition of the terms.

1.1 Background of the Problems

Test is an important part of teaching learning process that cannot be separated in the implementation of the teaching and learning process itself. Test begins with a close collaboration between curriculum experts and measurement experts from major universities, school districts, and test publishers. These experts identify major academic skills and bodies of knowledge that students are expected to know and then they create appropriate test questions to assess the students skills and knowledge.

In teaching learning process, testing is important to measure the ability of the students. Heaton (1990:5) states that, both teaching and testing are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. The objectives of learning can be evaluated by having a test, since the test is constructed to find out the achievement of the learners in teaching and learning process.

A test has the purpose of measuring the testee’s performance. It is intended to measure a student’s ability or knowledge. The result of the test is used to evaluate the progress of teaching and learning process, since the performances of the learners have to be evaluated progressively. Besides, the result of the test can be used to see how successful the teaching and learning process is implemented. As stated by Woods (2005:25), this is important as testing can often influence the nature of what is taught and how it is taught.

It is quite clear that by having a test, the teacher can see how far the students understand the materials that he or she delivers and how well the materials are delivered. Therefore, the good test will be able to provide the accurate information about learners' performance and the quality of the teaching and learning process.

To measure the students' ability, multiple choice testing can be an efficient and effective way. Consequently, multiple choice tests assumed as the most standardized test in testing the students' knowledge. Some people claim that multiple-choice tests can be useful for measuring whether students can analyze the material.

Multiple choice tests ask a student to recognize a correct answer among a set of options that include 3 or 4 wrong answers that is called distracters. The decision to use multiple-choice tests or include multiple-choice items in a test should be based on what the purpose of the test is and the uses that will be made of its results.

To achieve the purpose above, the teachers have to make sure that all the test items they have made are standardized. As the measurement to evaluate the students' understanding in the teaching and learning process, the test should meet the criteria of a good test. There are several aspects that constitute the criteria of a good test. According to Sulistyo (2007:21) who states that a good test must meet the requirements: reliability, validity, practicality/usability, and economy. If a test does not meet the requirements of a good test, the test will produce biased scores that will not reflect the real ability of the test takers. Then, it is obvious that the good test must meet those criteria so that the results of the test can be definitely relied on.

The test is a procedure that can be used to determine or measure something in accordance to the way and the rules that have been set (Suharsimi, 2003: 46), while the non-test is a procedure used to measure the affective domain such as attitudes, interests, talents , and motivation, for example using questionnaires, interviews, observations, and others (Sudijono, 2011: 67). Although there are two kinds of measuring instruments evaluation activities, but the test is often used for the evaluation tool. The test that is developed in this case is known as teacher-made test.

The teacher-made test here is constructed and administered by the teacher. Hence, the items of the test are not analyzed systematically by the teacher. Arikunto (2003:147) states that the teacher-made test is constructed from the items that are commonly not tried out, analyzed, and revised first. Therefore, based on that case, the quality of the teacher-made test is questionable.

Regarding the quality of the teacher-made test, it is obvious that the teacher-made test needs to be tried out and analyzed. The results of the analysis shows which items have good quality and which items need to be revised. Therefore, the teachers need to be capable to make a good test to help their students learn.

As far as the research is concerned, test should measures four aspects, i.e, validity, reliability, level of difficulty and discrimination power. However, most of the teachers are still using manual method to analyze the quality of their test items. The way of using manual method might consumes much times. Actually, the teachers can use iteman software program to analyze the quality of the test items they have made.

Consequently, iteman is very important for the teachers in administering the test items. Basically, it is a software that is 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, and the quality of the alternatives. Hence, this software program can help the teacher in administering the test items in easy way.

As iteman is considered useful, the teachers are more expected to have an involvement in assessing the multiple choice tests using the item analysis program. However, based on the researcher's pre-observation in SMAN 5 Bandar Lampung, it was found that most of the English teachers were unfamiliar with iteman software program. They had never learned how to use iteman before. They had lack of knowledge about how to analyze the test items, to decide the validity, reliability, discriminating power, and level of difficulty, especially by using iteman program.

Besides, most of students rarely got good scores because based on their opinion, the test items were too difficult. It might be consumed that the test items which is made by the teacher were not consider as a good test based on the criteria. They did not know how to measure the quality of the test items they have made before those items were used. Hence, most of the students did not get good scores on the exam because of its difficulties.

Based on the explanation above, the problem concerning with the analysis of test items is considerably need to be investigated because the test items which will be used for testing the students have to analyzed first. Therefore, the researcher is interested in analyzing the Final Semester test items made by the teacher using Iteman software, which will be used for the first semester of the first grade at SMAN 5 Bandar Lampung. The purpose of the item analysis is to determine whether the test items are good or not, based on the validity, reliability, discriminating power, and level of difficulty.

1.2 Identification of the Problems

According to the background of the problem above, the following problems can be found:

1. The English teachers at SMAN 5 Bandar Lampung are unfamiliar and rarely analyzed the test to determine the quality of test items.
2. The English Final Semester test items made by the teacher is not identified in term of validity, reliability, discrimination power, level of difficulty, and the quality of alternatives.
3. The quality of the test items which made by the teachers are questionable because they have lack of knowledge about how to analyze the test items using iteman before the test items were used.
4. Most of the students had bad scores because the test items were too difficult.

1.3 Limitations of the Problems

Based on the identification of the problems above, this research focus on the quality of test items for Final Semester made by the English teacher at SMAN 5 Bandar Lampung by using Iteman software program. Most of the English teachers did not know how to analyze the test items they have made before they were used for the students. Hence, the students were rarely got good scores.

Moreover, the qualities of test items were questionable. The test items for the students have to be analyzed first to know whether they belong to a good test or not, based on the criteria of a good test, i.e., validity, reliability, level of difficulty and discrimination power. To ease the teachers analyze the test items they have made, iteman software program might be used. However, this kind of software program can not be used for essay items. Therefore, this research will focus on analyzing the multiple choice test items only.

1.4 Research Questions

This research is aimed at evaluating final test made by English teachers to have the quality of the test and propose some revisions.

In relation to:

1. How is the quality of the English final test made by the teacher based on the criteria of a good test (i.e., validity, reliability, level of difficulty and discrimination power) by using iteman software program?
2. What revision should be made by the teacher on English final semester test items at SMAN 5 Bandar Lampung?

1.5 Objectives of the Research

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

1. To evaluate the quality of English final semester test items made by the teacher at SMAN 5 Bandar Lampung based on the criteria of a good test. Specifically, this research identified the validity, reliability, level of difficulty and the discrimination power of the test itself by using iteman.
2. To know what revisions which should be made by the teacher at English final semester test items at SMAN 5 Bandar Lampung.

1.6 Uses of the Research

The findings of this research are hopefully can be useful both theoretically and practically.

1. Theoretically, the results of this theory are expected to complete the previous theories of the quality of assessment.
2. Practically, this research may be used to help the teachers assess the quality of multiple choice tests by using iteman software program.

1.7 Definition of Terms

As a prevention of misunderstanding from the reader, the definition of terms which are used in this study are provided as follows:

1. Discriminating Power

   Discriminating power of test items refers to the percentage of high-scoring individuals responding correctly versus the number of low-scoring individuals responding correctly to an item. Or, in other words, it refers to the ability of the test items to discriminate between the clever and the lowest students.

2. Level of Difficulty

   Level of difficulty of test items is the percentage of students answering correctly each items in the test. this numeric index indicates how effectively an item differentiates between the students who did well and those who did poorly on the test.

3. Reliability

   Reliability of test items is the consistency of a measurement. A test is considered reliable if similar results are obtained repeteadly. For Example, if a test is designed to measure a trait, then each time the test is administered to a subject, the results should be approximately the same.

4. Validity

   Validity of test items is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order that the resuls can be accurately applied and interpreted.

5. Test Items

   Test items are questions relating to reading comprehension which consist of a stamp and for load by options (a,b,c and d) which ask a certain item such as identifying main idea, making inference, identifying synonym, antonym, identifying reference, and the like

II. LITERATURE REVIEW

This chapter discusses two major points: review of previous research and review of related literature. They are elaborated in the following sections.

2.1. Review of Previous Studies

In relation to this research, there have been several studies related to the quality of English test items in general. There are three researchers who have conducted research on the quality of English test items (Lestari, 2010; Putri, 2009; and Nurung, 2008).

Lestari (2010), investigated 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. From the finding, it can be seen that the quality of test items for multiple choice test are not good enough based on the criteria of a good test. It can be assumed that the teacher do not pay attention to the aspects while creating the test items.

Another research was done by Putri (2009). 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. 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.

Furthermore, Nurung (2008) 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 19 (47.5%), while the bad test items are 21 (52.5%). The percentage of the bad test items is higher than the good test items. Concerning the finding above, the revisions for the test items should be needed. Teachers should able to creates the good test items.

Based on the results of the previous studies, several findings have been identified. It can be stated that all of the previous studies implies the importance of analyzing the test items. The studies showed the criteria of a good test which must be applied for the teacher in creating the test items before it is used. Besides, the studies above help the researchers to build their ideas on evaluating the test items made by the teacher based on the quality of a good test.

However, there is still, at least one issue that has not been found, that is, how to evaluate the quality of English final-semester test items made by the teacher based on the criteria of a good test i.e., the validity, reliability, level of difficulty and discrimination power. Therefore, this study concerned about the quality of the final test based on the criteria of a good test such as validity, reliability, level of difficulty, discriminating power, proportion of the answers and distractors.

2.2. Review of Related Literature

For the specific explanation about evaluating final semester test using iteman software program, the researcher explains some related literature about concept of test, type of test, multiple choice test, quality of test items and iteman software program.

2.2.1 Concept of Test

A test is used to see whether or not the test actually tests what should be tested. Haladyna, (2004:4) states that a test contains a single item or set of test items intended to measure a domain of knowledge or skills or a cognitive ability. Tuckman, (1975:8) 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 is a process that must be done in teaching learning activity.

For measuring the students' knowledges, the procedure of a test should be systematic. Brown (2004:3) defines test as an instrument that provides an accurate measure of a person's ability, knowledge or performance in particular domain. Carrol (1968: 46) states that educational test as a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual.

Based on the definitions above, it can be concluded that a test is a necessarily instrument to measure student's outcomes as skill, ability, knowledge, talent, and the others. The students must give their best performance by answering a set of questions given by the teacher in the class to represent their understanding. It should reflect the knowledge that has been taught.

2.2.2 Type of Test

Based on the function in learning activity, Purwanto, (2011: 67-70) classified the test into four kinds, they are formative test, summative test, diagnostic test, and placement test which has comprehended by student.

a. Formative Test

Formative test is intended as test which used to determine student's comprehension after learning activity. Every learning program or learning material must comprehend the student's in line with the learning purpose which has arranged. Formative test is examined to determine the effectivity of teaching learning process which has comprehended by student.

Formative test in learning practice called as ulangan harian or daily test. In teaching planning, learning process in a sub material is planned in a lesson plan. Lesson plan covers learning purpose, material, method, learning strategy, media and evaluation. Evaluation which is planned in a lesson plan is evaluation that is done based on formative test. Formative test given in the end of learning program, so it could be called as post test.

b. Summative Test

Summative test is intended as a test used to determine student's comprehension toward all of materials which have given in a certain period time such as mid semester and semester. In teaching practice, summative test known as mid-term test and final test semester.

c. Diagnostic Test

Learning outcomes evaluation has diagnostic function. Test which has used to do diagnostic evaluation is diagnostic test. Diagnostic test used to identify student's difficulty and investigate kind of difficulty which is faced. Based on the student's comprehension and difficulty faced, teacher could solve the student's difficulty with appropriate solution. Example of diagnostic test is to teach descriptive test, teacher should be sure if student comprehends about simple present tense. Before starting about descriptive test, teacher gives diagnostic test to determine student's comprehension about present tense.

d. Placement Test

Placement test is intended to collect the data about learning outcomes test which is needed to grouping student with their interest and their talent appropriately. This grouping is done in order to give appropriate materials with their interest and talent. In learning practice, placement test used for example in Senior High School to grouping students in IPA, IPS or Bahasa. Placement test serves data to put students in a class appropriate with their interest and talent.

While in this research, the test which will be analyzed is summative test, in case of final semester test. It will measure the students' comprehension toward all the materials given by the teacher in a certain period time. Moreover, the teachers' ability in creating test items based on the quality of a good test will be investigated.

2.2.3 Multiple Choice Tests

Multiple choice test consits of an information about an incomplete definition. Gronlund (2009) stated that multiple choice consists of a problem and a list of suggested solutions. To complete the definition, student should choose one from some possible answers. Multiple choice consists of information and the possible answers. The problem may be stated in form of direct question or an incomplete statement or it is called stem. The lists of suggested solution are called alternative. The correct alternative in each item is called answer and the remain alternative is called distractors. The possible answer covers one correct answer and two until four distractors. It is recommended by using four alternatives for grammar items, but five for reading and vocabulary.

The alternatives in multiple choice test 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. While the best-answer type of multiple choice is useful for measuring and learning outcomes that require understanding, application, or interpretation of factual information. In spite of it is easy to score, these types of test are difficult to be constructed and give the students a possibility to guess.

Heaton (1990:27) shows the positive effect of using multiple choice in measuring learning outcome. It can measure various learning outcomes, from simple to complex. Some advantages of multiple choice item are good for testing knowledge of grammar, vocabulary, rather than the ability, measuring students' ability to recognize grammatical form. Not only advantages, multiple choice item has negative criticism too. It is constructing negative characteristic especially reduce students knowledge and creativity.

2.2.4 Guidelines for Constructing Multiple Choice Test Items

According to Gronlund (1967) 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.

Isaacs (1984) states that 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 orby 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.

Based on the explanation above, all of the steps are important and good for test maker or the teachers to make the good and qualified test items.

2.2.5 Guidelines for Constructing Distractors

When test writers refer to style, they usually mean the expression of ideas in a smooth, orderly, pleasing manner. Each test writer develops an individual style of expression that allows for a personal presentation of his or her own thoughts and emotions. Suparman (2011) stated that when a multiple choice item is going to be constructed, there are six principle guidelines to be followed by a test developer as follow:

1. Each multiple choice item should have only one answer. This answer must be absolutely right if the instruction does not specify choosing the best option. Although this may seem an easy meter, it is sometimes very difficult to construct an item having a\one correct answer.
2. Only one feature at time shoukd be tested. It has long been standard practice to test only onw feature at time. it is usually less confusing for the tested and it helps to reinforce a special teaching point. Clearly, few would wish to test both grammar and vocabulary at the same time, but sometimes would order and sequence of tenses are tested simultaneously. Such item is called impure item.

3. Each option should be grammaticaly correct when placed in stem, except in the case of specific grammar test item.
4. All multiple choice test item should be at a level appropriate to the linguistic ability of the testees. The context itself shoukd be a lower level than the actual problem which the item is testing.
5. Multiple choice item should be as brief and clear as possible.
6. In many tests, items are generally arranged in rough order of increasing difficulty. It is generally considered important to have one or two simple item to lead in the testees, particularly if they are not familiar with the kind of test being administered.

Well-constructed Multiple Choice questions are time consuming and difficult to write. Furthermore, one aspect where many of Multiple Choice question fail is in having effective distractors. Teachers often spend a great deal of time constructing the stem and much less time on developing plausible options to the correct answer. High quality of Multiple Choice question, however, also need the options to be well written. In a classroom setting where test items are designed to measure educational outcomes, distractors must perform acceptably and each distractor should be based on a common misconception about the correct answer. Millman and Greene (1993) states that a discrimination index or discrimination coefficient should be obtained for each option in order to determine each distractor's usefulness. The purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item. Conversely, the distractors must appear as implausible solutions for those students who have achieved the objective. Only the answer should appear plausible to these students.

2.2.6. Quality of Test Item

Test is commonly used to assess the student's knowledge and the outcome of the learning process. The test should be qualified and reflect the materials that have been taught before in order to get the good score of the students. According to Athiyah (2012) a test can be said as a good test if it fulfills several requirements of a good test, both statistically and non statistically.

a. Validity

Validity is one of the important aspects of the test. The test will gain nothing if the validity is not valid. 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). The test can be considered as good test if it can really measures the quality of test.

To measure whether the test has good quality, the test should have face validity, content validity, and construct validity.

1) 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. However, face validity is assumed as not as important compare with other indications of validity.

2) Content Validity

   Content validity represents the correlation between the test and exact materials, in terms of construction. It is concerns with whether the test is sufficiently representative and comprehensive. Shohamy (1985:74) defines 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 wants the students to know. Content validity is important because it would give the information whether the students understand the material or not. It means, the test items of the test should present the material being discussed.

3) Construct Validity

A test can be considered to be valid if the item of the test can measure every aspects which is suitable with the specific objective of the instruction. Construct validity will be concern 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 final semester test items which is made by the English teacher should really measure the students' understanding. Thus, the test can be said to be construct valid if it can measures the construct or theoretical ideas.

b. 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 (Linn & Gronlund, 2000:193). If the results of a test are replicated consistently, they are reliable. Test reliability is important because it is necessary for a good validity. A test can be highly reliable without necessary being valid for any purpose of interest. Test reliability is refers to the reproduced ability of test results. In short, 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 conditions or situations.

c. Discriminating Power

Discrimination power is an aspect of item analysis, discrimination power tells about which item discriminates between the good and not good 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.

d. Level of Difficulty

Level of difficulty is the percentage of correct answer from the students who take the test. 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. Level of difficulty concerns with how difficult or easy the items for the students. If the students's answers are mostly correct, it means that the test items are too easy.

The difficulty level of an item is known as index of difficulty. Index of difficulty is the percentage of the students who answer correctly on each test items. Index of discrimination refers to the percentage of high-scoring individuals responding correctly versus the number of low-scoring individuals responding correctly to an item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). This numeric index indicates that an item can be effectively differentiates the students' who did well on the test and those who did not.

An item is considered good if the difficulty index is 50%. It can be said that the item is neither easy nor difficult. If an item has a difficulty index of 67.5%, it indicates that 67.5% easy and 32.5% difficult. The information of the difficulty index of an item can help the teacher to decide whether a test should be revised, retained of modified.

It is necessary to analyze the quality of the test items before it is given to the students. According to Arikunto (2006), item analysis is a systematic procedure, which will provide information that is very specific to the test items arranged. In iteman software program, the measurement of validity is not covered explicity. In order 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). Therefore, the conclusion from the three aspects gives a decision whether the test item has good validity or not.

2.2.7. Iteman Software Program

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 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 teachers' skills in test construction, and identifying specific areas of course content which need the students emphasis or clarity. 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 according to Linn & Grondlund, (2000). 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 Linn & Grondlund (2000) mentions the importance of item analysis.

There are:

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.

2.2.8. 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.

The advantages of Iteman in assessing the students ability are as follows:

1. Iteman is a simple application. Iteman is easy to use and very simple, the researcher just need an electricity, computer and Iteman software program to analyze the data.
2. Iteman can be used everywhere, anywhere and for everyone. Iteman easy to understand. Everyone can use Iteman, because steps to use Iteman is very easy and simple. The researcher needs 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 the researcher input the data to the computer, then it 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 beused.
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.

III. RESEARCH METHOD

This chapter concerns about the methods of the research used in this study, which include research design, population and sample, data collecting technique, research procedures, and criteria of a good test, and data analysis.

3.1 Research Design

The design of this current research was descriptive and evaluative which described the result of an evaluation on an object based on standard criteria using iteman. This research was intended to evaluate and propose some revisions for final semester test in the first grade in SMAN5 Bandar Lampung. The objects of this research consisted of test items and the student's answer sheets. Both of them were analyzed based on standard criteria, that is, level of difficulty, discriminating power, reliability and validity.

3.2 Setting

This research was conducted at the first year of SMAN 5 Bandar Lampung. The research conducted in a week. It was administered during the English lesson which was being tested when the students had finish their English final semester test items.

3.3 Object of the Research

The object of this current research was teacher-made English final semester test items for the first grade on the first semester at SMAN 5 Bandar Lampung. The number of answer sheets which was used in this research was around 30. The test items made by the teacher was tested to get the data of the student's answers.

3.4 Data Collecting Technique

Final semester test items were collected as the data for this research. The test was tested to determine whether there will be some revisions for the test items based on the result of analysis while using iteman.

The data was collected by administering the teacher-made English final-semester test items to the first year student of SMAN 5 Bandar Lampung. There were 45 questions. The data consisted of students' answers which put on the answer sheets.

3.5 Research Instrument

There was an instrument to gather the data of the students' answers, that was a teacher-made English final-semester test items as a document in the first grade of SMAN 5 Bandar Lampung. This document was obtained by approaching the headmaster and the English teacher. During conducting the approachment, the researcher first asked the headmaster to permit her to carry out the research in the school, and then asked him to give the permission to access the document for the research.

3.6. Research Procedures

To check the quality of the final semester test, there were several procedures to obtain the question sheets and the students' answers. The instrument was the final semester test; each item has five options, they are A, B, C, D and E. Then, the researcher will analyzed the test.

To make the research run well, there were several procedures 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 was all of the final semester test items at the first grade of SMAN 5 Bandar Lampung.

3. Determining the class

   The researcher took one class. The sample of this research was X SOS 4 class.

4. Determining the test

   The test was from the final examination. There were at least fourty multiple choice test items.

5. Carry out the test

   The students were given the test by the teacher and they should answer the questions. The allocation time was around 60 minutes.

6. Collecting the students' answers

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

7. Analyzing the data quantitatively

   This research touched the final semester test by counting on Iteman software program.

8. Analyzing the data descriptively

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

To analyze the data using iteman program, all the data must be put and saved in Notepad. According to Suparman, (2011), the following are the steps of utilizing the program:

1. Open iteman program by clicking start
2. Select program/ click iteman

3. Type the name of your data file (input) as you like on Enter the 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 will appear. 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: For example, D:\Midtest.scr, then click Enter. Finish.

3.7. Data analysis

The data which were collected by the researcher by means of administering the tests that the teacher made for final semester had been analyzed. The test was administered by the researcher together with the English teacher. The data analysis focused on evaluating the test items to find out whether the test items are good or not, seen from the points of: validity, reliability, discriminating power, and level of difficulty.

Whereas, the data analysis was also intended to determine the interpretation of each item, that is, whether each item can be used well, should be totally revised, or partially revised, or dropped totally.

To interpret the results of analysis test items, the researcher used the criteria of the quality of test items by some experts in Suparman (2011):

Table 3.1 Criteria of Test Item Quality

Prop Correct (Level of Difficulty – p)	
0.000 – 0.250	Difficult
0.251 – 0.750	Average
0.751 – 1.000	Easy

Point Biserial (Discriminating Power – D)	
D ≤ 0.199	Very low
0.200 – 0.299	Low
0.300 – 0.399	Average
D ≥ 0.400	High
Prop Endorsing (Proportion of the Answer)	
0.000 – 0.010	Low
0.011 – 0.050	Sufficient
0.051 – 1.000	Good
Alpha (Test Item Reliability)	
0.000 – 0.400	Low
0.401 – 0.700	Average
0.071 – 1.000	High

Source: Suparman (2011: 95)

Furthermore, to make the teacher or the assessor easier in choosing the test items which need to be revised or dropped, the following guideline can be considered as the reference:

Table 3.2 Criteria to classify the quality of test items

Prop Correct (Level of Difficulty – p)	
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
0.091 – 1.000	Very easy/needs dropping or total revising
Point Biserial (Discriminating Power – D)	
D ≤ 0.199	Very low/needs dropping or 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 (Proportion of the Answer)	
0.000 – 0.010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very good
Alpha (Test Item Reliability)	
0.000 – 0.400	Low/not sufficient
0.401 – 0.700	Average/sufficient
0.071 – 1.000	High/good

Source: Suparman (2011: 95-96)

V. CONCLUSIONS AND SUGGESTIONS

This chapter deals with two major points, those are conclusions and suggestions based on the results and the discussions of this research, and elaborated in the following sections.

5.1. Conclusions

The findings of this research showed that not all items in the English final semester test have high reliability, average level of difficulty, high discriminating power, and good proportion endorsing. Meanwhile, some of the items had low reliability, low level of difficulty, low discriminating power and least proportion endorsing. It means that there were some items which were not standardized to test the students and need to be revised to make the test items properly.

Based on the result in the output data in the iteman, the following conclusions are drawn as follows:

1. The validity of the English final test semester test items was tried out, analyzed and revised. However, the validity of the test items should be compared with the current of English curriculum used in SMAN 5 Bandar Lampung. Based on the discussion between the researcher and the English teachers at SMAN 5 Bandar Lampung, each of the items was relevant with the syllabus and some items were not. Concerning with the face, content and construct validity, the test items were not fulfill the requirements. Therefore, the test which is prepared by the English teacher is considered to be not fully valid.

2. The reliability of the English final test semester test items based on the result of the iteman was categorized as average/sufficient because the alpha value (reliability of the test items) was 0.492 which lied between the ranges of 0.401-0.700. It means that the reliability of this English final semest test items was categorized as average/good. the test items were proper enough to be tested to the students because they are reliable
3. The level of difficulty of the English final semester test items were classified into five categories, i.e., average/good, very difficult, difficult, easy, and very easy. Based on the Proportion Correct (level of difficulty), there were 16 out of 45 (35.56%) test items which considered good or average. The question can be said not good or has low quality because most of the items included in the category the items should be dropped or revised. Problem with these categories can be repaired by replacing the question where some students were able to answer it because it is likely most of the students had comprehended the material in the questions.
4. The discriminating power of the English final test items can be classified into four categories, i.e., high, quite average, low/need revising, very low. It indicated that some of the items fulfilled the requirements of the quality of the good test item but some of them did not. There were more than 60% items should be dropped, meanwhile only around 20% items that can be used directly without any revision. This suggested that the teacher should revise many items before using them
5. Proportion endorsing (the qualities of the options) in English semester test items, regarding on the iteman analysis were classified into three classifications, i.e., least/drop, good enough/sufficient, and very good. It was obtained that the options of the 45 items each of which consist of A, B, C, D, and E totaling 225 options. From the result of analysis by using

iteman, it was found that 95 options should be revised because they were classified into low category. It can be concluded that not all the test items have good quality and can be accepted. Less than 50% the items should be revised by the teacher before giving the test item to the student in the examination.

6. Regarding to the interview with the English teachers, it can be concluded that they never analyse the test items before it was used to the students. The teachers had not been familiar with the iteman, hence they were not using it to determine whether the test items they had made were propoer enough to used or not.

5.2. Suggestions

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

1. The teachers should be able to make a proper test items for the students based on the quality of a good test before it is used.
2. The teachers should be familiar with all the terms related to the quality of a good test, such as validity, reliability, prop. Correct (level of difficulty), point biserial (discriminating power), prop. Endorsing (the alternatives/ options), distracters, key answers, alpha and standard deviation.
3. The teachers should be familiar with the iteman program to make them easier in assessing the student's ability.
4. The teachers should be trained to use the item analysis program (iteman) in order to improve the quality of the test.
5. The test items which is made by the teachers should be tried out first, before it is used to the students.
6. The teachers should be trained on how to analyze the test items effectively and efficiently and how to revise the bad test items.

7. The researcher should be able to analyze the other test items, such as Mid Semester test, Final School test (UAS), and National Examination (UN)

REFERENCES

Anderson, J., & Hughes, A. (1981). *Issues in Language Testing.* London: The British Council.

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

Arikunto, S. (2003). *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi).* Jakarta: Bhumi Aksara.

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

Boopathiraj, & chellamani. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in education. *International Journal of Social Science & Interdisciplinary Research*, 189-190.

Carrol, J. (1968). *The psychology of language testing. In A. Davies (Ed.). Language Testing Symposium.* London: Oxford University Pressed.

Corporation, A. S. (1989-2006). *User's Manual for the ITEMAN Conventional System Analysis Program.* . Suite 200: 2233 University Avenue.

Cronbach, L., & Meehl, P. (1995). Construct validity in psychological. *Psychological Bulletin*, 281-302.

Gronlund, N. E. (1967). *Measurement and evaluation in teaching.* New York: Mac. Milan Co.

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

Hatch, E., & Farhady, H. (1982). *Research Design and statistics for applied linguistics.* Rowley, Massachusetts: Newburry House.

Heaton, J. B. (1990). *Writing English Language Tests.* New York: Longman inc.

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

Isaacs, G. (1994). *Multiple choice testing: A guide to the writing of multiple choice.* Campbelltown: NSW: HEROSA.

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.

Linn, R., & Gronlund, N. (2000). *Measurement and evaluation in teaching (8th ed.).* New York: Macmillan.

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

Millman, J., & Greene, J. (1993). *The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), Educational measurement.* Phoenix, AZ: Oryx Press.

Nurung, M. (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.* Yogyakarta: Graduate School, State University of Yogyakarta.

Osterlind, S. (1998). *Constructing test items# &ultiple'choice, constructed'response, performance and other formats.* New York: Kluwer Academic Publishers.

Purwanto. (2011). *Evaluasi Hasil Belajar.* Yogyakarta: Pustaka Belajar.

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.* Undergraduates thesis: Universitas Negeri Semarang.

Rajhy, A. (2014). FIVE CHARACTERISTICS OF A GOOD LANGUAGE TEST. *National Journal of Extensive Education and. Interdisciplinary Research*, 2-6.

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

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, Vol. XII. No. 1.*, Hal. 1-96. ISSN 1411-2501.

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

Wood, D. (1960). *Test construction: Development and interpretation of achievement tests.* Columbus, OH: Charles E. Merrill Books, Inc.

Wood, D. (1960). *Test construction: Development and interpretation of achievement tests.* Columbus OH: Charles E. Merrill Books, Inc.

Woods, C. (2005). *Professional Development for Teachers: Teaching and Assessing Skills.* Cambridge: Cambridge University Press.