

CHAPTER 2

THEORITICAL 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: Ariyana, 2011; Fitriana, 2013; and Ratnaningsih, 2009.

Ariyana (2011) conducted research on students of SMP in Grobogan district, Semarang. She investigated final semester test at the third semester in science class. The purpose of the research was to find out the validity, level of difficulty, discriminating power, the effectiveness of alternatives, and the reliability. The method of collecting the data was recording. The quantitative approach was done by using ITEMAN program. The result of the research showed that the multiple choice test was 2% very difficult; 20% difficult; 70% average; 4% easy and 4% very easy. The discriminating power of the test was that 26% was average, 62% was high, 10% was low, and 2% was very low. The effectiveness of the alternatives was 82%. It meant that the alternatives were functional. The reliability of the test was 0.711 which

meant that the test had high reliability. Based on the result of the research, the validity was reasonable but needed revising. Therefore, the multiple choice test had average level of difficulty, high discriminating power, functional alternatives, and high reliability.

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 P 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.

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 used ITEMAN as a tool to analyze multiple choice tests in elementary school, junior high school, and university as the population and sample of the research. As known that, the validity in the ITEMAN is concluded by covering level of difficulty, discriminating power, and proportion of the distracters. This research discerns the validity seen from the content validity, construct validity, and face validity. Because of that, the researcher analyzed on those sides and investigated the population which had different knowledgeable students and multiple choice items, as the focus of this research.

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.

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. Nunnally (1978:301) states that item analysis is extremely useful. This furnishes a variety of statistical data regarding how subjects responded to each item and how each item relates to overall performance. From the two definitions above, it can be concluded that the analysis of the test is a systematic activity that involves the collection and processing of data in the form of a test that is done in order to obtain information to determine a conclusion about the quality of the test.

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 approach is 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. In determining the characteristics of the item, there are generally three things which should be considered, namely: (1) level of difficulty, (2) discriminating power, and (3) effectiveness of distracters. These three characteristics of the item jointly determine the quality of the item. Linn & Gronlund (1995:47) define that a good test must have three characteristics, namely validity,

reliability, and usability. Validity means that the accuracy of the interpretation of the results of the measurement. Reliability means the consistency of the result measurement, and usability means the procedure is practical.

a. Validity

If the result of a test is not considered valid, then the test is meaningless. If it does not measure what it is measured, the result cannot be used to answer the research question, which is the main aim of the research. Validity is the extent to which an instrument measures what it is supposed to measure (Carmines and Zeller, 1979:17). According to Lynne (2004:31), validity, reliability's partner term, refers to the ability of the test to measure what it is required to measure, its suitability to the task at hand. Besides, according to Wiggins & McTighe (2005:194), validity refers to the meaning the raters can and cannot properly make of specific evidence, including traditional test-related evidence. For the criteria of validity, in a very general sense, a test is valid for anything with which it correlates (O'neil, 2009:23). Therefore, validity almost seems like an afterthought, in some ways drawing upon the overall history of validity in which the test authors are the supreme authority about the validity of their tests.

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 Listening

Listening is one of the most fundamental skills in learning language. Because a communication will not be running well if this basic skill is not mastered, especially for ESL. Listening is an activity of paying attention and trying to get meaning from something through ears. In listening comprehension, the forms of the test that are given to the testees are short utterances, dialogues, talks, and lectures (Heaton, 1975:8). It indicates that the listener must digest the message of the speaker carefully to get the information from the speaker. For listening

comprehension, he defines that an effective way of developing the listening skill is through provision of carefully selected practice material. Such material is in many ways similar to that used for testing listening comprehension. He considers that it is possible to develop listening ability if the practice material is not dependent on spoken responses and written exercises.

Based on the statements above, listening is a manner conducted by the listener in actively paying attention and understanding the meaning of the words the speaker says.

b. Traits of Speaking

Speaking is an action of conveying ideas and thoughts. It takes the part of pronunciation, vocabulary, grammar, fluency and comprehension altogether (Haris, 1974:84). According to Heaton (1975:8), to test speaking ability, the test is usually in the form of an interview, a picture description, role play, and a problem-solving task involving pair work and group work. Therefore, speaking test can take place if the speaker uses verbal symbol like word and non verbal symbol like gesture and body language to convey the intention.

c. 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.

d. Traits of Writing

Writing is a productive skill in the written form. Writing is one of the language skills that are used for indirect communication such as, letter, note, short message, and invitation. Through writing, students can express their understanding of problems or ideas. Writing is considered the most difficult skill to master (Shohamy, 1985:188). Moreover, Heaton (1975:135) says that this skill needs not only grammatical and rhetorical devices, but also conceptual and judgmental elements. Writing is a productive skill in the written form. According to Heaton (1974:135), five components that are necessary for testing the writing skills are:

1. Language use: the ability to write correct and appropriate sentences.
2. Mechanical skills: the ability to use correctly those conventions peculiar to the written language – e.g. punctuation, spelling.
3. Treatment of content: the ability to think creatively and develop thoughts, excluding all irrelevant information.
4. Stylish skills: the ability to manipulate sentences and paragraphs, and use language effectively.

5. Judgment skills: the ability to write in an appropriate manner for a particular purpose with a particular audience in mind, together with an ability to select, organize and order relevant information.

e. 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 DeCapua (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.

f. 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.

b. Reliability

If the results of a test are replicated consistently, they are reliable. In psychometrics, reliability is a technical measure of consistency (Lynne, 2004:31). Reliability is the degree to which a test consistently measures whatever it measures (Crocker & Algina, 1986:105). Therefore, any random influence which tends to make measurements different from occasion to occasion or circumstance to circumstance is a source of measurement error (Nunnally, 1978:248). In ITEMAN software program, Alpha is the measurement of reliability of a test.

There are three indexes that can be followed to determine whether the reliability of a test is very bad, sufficient, and very good, as follows:

Table 2.1 Criteria of Reliability (Alpha)

| Criteria | Index | Classification | Decision |
|---------------------|---------------|----------------|------------|
| Reliability (Alpha) | 0,000 - 0,400 | Low | Very bad |
| | 0,401 - 0,700 | Average | Sufficient |
| | 0,701 - 1,000 | High | Very good |

Source: Ngadimun (2004:8)

c. Usability

A test is said to have a high usability when the test is practical. That is, the test is easy to be implemented, easy to be assessed, easy to make administration, and also fulfilled with clear and complete instructions that may be given by others.

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.

In fact, traditional assessment is still implemented and used in final semester test. Multiple choice tests are the test which is still counted on by MGMP. This type of assessment is not the only way or the best way to evaluate students, but is the most common way used to measure the student learning process.

2.2.3. Multiple Choice Tests

This kind of test requires the students to pick out the correct answer from several alternatives provided by the test maker. 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 indirect measures of performance. A standard multiple choice test item consists of two basic parts: a problem (stem) and a list of suggested solutions (alternatives). The stem may be in the form of either a question or an

incomplete statement, and the list of alternatives contains one correct or best alternative (answer) and a number of incorrect or inferior alternatives (distracters) (Crocker & Algina, 1986:76). 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).

2.2.4. Guidelines for Constructing Multiple Choice Items

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. For analyst, however, style connotes something different. Editorial style refers to the consistent use of a set of rules and guidelines. The rules and guidelines prescribe a consistent use of punctuation, abbreviations, and citations, a uniform and attractive format for tables, graphs, and charts, and a correct form for the many other elements that constitute written communication (Osterlind, 1998:161).

There was one research by Haladyna and Downing (1989a, 1989b) in Haladyna (2004:98) involving an analysis of 46 textbooks and other sources on how to write multiple choice test items. The result was a list of 43 item-writing guidelines.

Author consensus existed for many of these guidelines. But for other guidelines, a lack of a consensus was evident. The next study by them involved an analysis of more than 90 research studies on the validity of the item-writing guidelines. Only a few guidelines received extensive study. Nearly half of the 43 guidelines received no study at all. Since the appearance of these two studies and the 43 guidelines, Haladyna repeated this study. They examined 27 new textbooks and more than 27 new studies of the guidelines. From this review, the guidelines were reduced to be 31 guidelines, which were used in this research. In such manner, he stated that there are two categories of item whether the item correlates to the guidelines or not, that is, flawed and non-flawed items. Because of that, these guidelines help the researcher determine the validity of the final semester test, especially in terms of face validity.

This research has a set of multiple choice item-writing guidelines that apply to all multiple choice formats taken from Haladyna's item-writing guidelines. So, the researcher implements the guidelines judiciously but not rigidly in determining how the face validity of the final semester test is.

2.2.5. ITEMAN Software Program

The use of ITEMAN stays widespread, but, some takes into account of an out dated system. 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.

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 except validity.

Proportion Correct

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). This value is generally reported as a proportion (rather than percentage), ranging from 0.0 to 1.0. A value of 0.0 would indicate that no one answered the item correctly. A value of 1.0 would indicate that everyone answered the item correctly.

There are four indexes that can be followed to determine whether a test item is rejected, revised, or accepted, as follows:

Table 2.2 Criteria of Proportion Correct (p)

| Criteria | Index | Classification | Decision |
|-----------------------------|---------------|-----------------------|-------------------------|
| Proportion Correct (p) | 0,000 - 0,099 | Very difficult | Rejected/total revising |
| | 0,100 - 0,299 | Difficult | Revised |
| | 0,300 - 0,700 | Average | Good |
| | 0,701 - 0,900 | Easy | Revised |
| | 0,901 - 1,000 | Very easy | Rejected/total revising |

Source: Ngadimun (2004:8)

Discrimination Index

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). This is also known as Differentiation Index. This statistic is a measure of each test question's ability to differentiate between high scoring and low scoring students. This is computed as: the number of people with highest test scores (top 27%) answering the item correctly minus the number of people with lowest scores (bottom 27%) answering the item correctly, divided by the number of people in the largest of the two groups.

$$\text{Disc. Index} = P_{\text{High}} - P_{\text{Low}}$$

Where P_{High} is the proportion of examinees in the upper 27% of the score distribution answering the item with the correct/keyed answer and P_{Low} is the same proportion in the lower 27% group.

The higher the number, the more the question is able to discriminate the higher scoring people from the lower scoring people. Possible values range from -1.0 to 1.0. A score of -1.0 indicates that the lowest 27% of the group all answered the

question correctly, and the upper 27% of the group all answered the question incorrectly. A score of 1.0 indicates that the upper 27% of the group all answered the question correctly and the lowest 27% of the group answered the question incorrectly. Negative discrimination would signal a possible key error (Haladyna, 2004:228).

There are four indexes that can be followed to determine whether a test item is rejected, revised, or accepted, as follows:

Table 2.3 Criteria of Discrimination (D)

| Criteria | Index | Classification | Decision |
|---------------------|----------------|----------------|-------------------------|
| Discrimination(D) | $D \leq 0,199$ | Very low | Rejected/total revising |
| | 0,200 - 0,299 | Low | Revised |
| | 0,300 - 0,399 | Average | Accepted |
| | $D \geq 0,400$ | High | Accepted |

Source: Ngadimun (2004:8)

Item-Total Correlation

This is recognized as correlation coefficients. These two coefficients are also known as Discrimination Coefficients (ASC, 1989-2006:13).

1. Biserial Correlation Coefficient

It is closely related to the point biserial correlation, with an important difference. The distinction between these two measures exists in the assumptions. Whereas the point-biserial statistic presumes that one of the two variables being correlated is a true dichotomy, the biserial correlation coefficient assumes that both variables are inherently continuous. Further, the assumption is made that the distribution of scores

for both variables is normal (Osterlind, 1998:282). For computational purposes, however, one of the variables has been arbitrarily divided into two groups, one low and the other high. In item analysis, the two groups are examinees who responded correctly to a given item and those who did not. In other words, it is a measurement of how getting a particular question correct correlates to a high score (or passing grade) on the test. Possible values range from -1.0 to 1.0. A score of -1.0 would indicate that all those who answered the question correctly scored poorly on (or failed) the test. A score of 1.0 would indicate that those who answered the question correctly scored well on (or passed) the test.

2. **Point Biserial Correlation Coefficient**

One index of discrimination is the point-biserial correlation coefficient. As a measure of correlation, the point-biserial coefficient estimates the degree of association between two variables: a single test item and a total test score (Haladyna, 2004:211). This statistic is a measure of the capacity of a test item (question) to discriminate between high and low scores. In other words, it is how much predictive power an item has on overall test performance. Possible values range from -1.0 to 1.0 (the maximum value can never reach 1.0, and the minimum can never reach -1.0). A value of 0.6 would indicate the question has a good predictive power, i.e., those who answered the item correctly received a higher average grade compared to those who answered the item

incorrectly. A value of -0.6 would indicate the question has a poor predictive power, i.e., those who answered the item incorrectly received a higher average grade compared to those who answered the item correctly.

The following 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. **Variance.** The variance of the distribution of examinee scores on each scale.
5. **Std. Dev.** The standard deviation of the distribution of examinee scores for each scale.
6. **Skew.** The skewness of the distribution of examinee scores for each scale.

The skewness gives an indication of the shape of the score distribution. A negative skewness indicates that there is a relative abundance of scores at the high end of the scale distribution. A positive skewness means that there is a relative abundance of scores at the low end of the distribution. A skewness of zero means that the scores are symmetrically distributed about the mean.

7. **Kurtosis.** The kurtosis of the distribution of examinee scores for each scale.

The kurtosis indicates the peakedness/flatness of the distribution relative to

that of a normal distribution. A positive value indicates a more peaked distribution; a negative value indicates a flatter distribution. The kurtosis of a normal distribution is zero.

8. **Minimum.** The lowest score on each scale for any examinee.
9. **Maximum.** The highest score on each scale for any examinee.
10. **Median.** The examinee score at the fiftieth percentile for each scale. It is thus the score that half of the examinees scored at or below.

There were 32 examinees in the data file.

Scale Statistics

```

-----
Scale:           0
-----
N of Items           30
N of Examinees      32
Mean                21.906
Variance             6.085
Std. Dev.           2.467
Skew                -1.504
Kurtosis             3.420
Minimum             13.000
Maximum             25.000
Median              22.000
Alpha                0.476
SEM                 1.786
Mean P              0.730
Mean Item-Tot.     0.294
Mean Biserial       0.445
Max Score (Low)     21
N (Low Group)       12
Min Score (High)   24
N (High Group)     10

```

11. **Alpha.** It is an index of the homogeneity of each scale. It can range in value from 0.0 to 1.0. This statistic is only appropriate for non-speeded scales

designed to measure a single trait. The alpha value is usually considered to be a lower-bound estimate of the reliability of a scale.

12. **SEM.** The standard error of measurement for each scale. It is an estimate of the standard deviation of the errors of measurement in the scale scores.
13. **Mean P.** The average proportion correct across all items on the scale for scales composed of dichotomously scored items.
14. **Mean Item-Tot.** The average point-biserial correlation across all the items in the scale.
15. **Mean Biserial.** The average biserial correlation across all of the items on the scale.

Listed above, these statistical measurements are the most widely used terms to assess multiple choice questions. The purpose of these reports is to help evaluate the quality of test items, and tests as a whole, by examining their psychometric characteristics.

2.2.6. 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 scored or multipoint items. The program can work only with multiple choice items. It is relatively easy to analyze test items using the ITEMAN program.

2.2.7. The Hypotheses

Based on the theories, the researcher formulated the hypotheses as follows:

- H₀ : The final semester test has not fulfilled the criteria of a good test, that is, has bad validity, low reliability, very easy or difficult level of difficulty, very low discriminating power, and non-functional alternatives.
- H₁ : The final semester test has fulfilled the criteria of a good test, that is, has good validity, high reliability, average level of difficulty, high discriminating power, and functional alternatives.