AN ANALYSIS OF THE TEST ITEMS OF ENGLISH FINAL EXAMINATION FOR THE SIXTH YEAR STUDENTS OF ELEMENTARY SCHOOL IN SOUTH SEMARANG REGENCY IN THE ACADEMIC YEAR OF 2007/2008

A FINAL PROJECT
Submitted in partial fulfillment of the requirements for the degree of
Sarjana Pendidikan In English

HARIS RIZQI ARIFIN
2201405671

ENGLISH DEPARTMENT
FACULTY OF LANGUAGES AND ARTS
SEMARANG STATE UNIVERSITY
2009

APPROVAL

The final project was approved by the Board of Examiners of the English Department of Languages and Arts Faculty of Semarang State University on September , 2009.

Board of Examiners

1. Chairperson

   Dra. Malarsih, M.Sn ----------------
   NIP.196106171988032001

2. Secretary

   Drs. Suprapto, M.Hum ----------------
   NIP. 195311291982031002

3. First Examiner

   Drs. Amir Sisbiyanto, M.Hum ----------------
   NIP. 195407281983031002

4. Second Advisor as Second Examiner

   Dr. Dwi Anggani L.B., M.Pd ----------------
   NIP. 195901141989012001

5. First Advisor as Third Examiner

   Drs. Hartoyo, M.A., Ph.D ----------------
   NIP. 196502231990021001

Approved by
Dean of Languages and Arts Faculty

Prof. Dr. Rustono, M.Hum.
NIP. 195801271983031003

DEDICATION

To

My beloved parents

My beloved lecturers

My beloved friends

and

My beloved girl friend

MOTTO

Success is my right. *(Andrie Wongso)*

Today is better than yesterday, and tomorrow must be better than today.

ACKNOWLEDGEMENTS

First and foremost the writer would like to express his gratitude to Allah SWT for his blessing and inspiration give to him in completing his study.

The deepest gratitude and appreciation is extended to Drs. Hartoyo, M.A, PhD, his first advisor, for his patience to guide, advice and encourage him since the beginning until the end of this final project writing has been completed. Furthermore, he would like to express his deepest gratitude to Dr. Dwi Anggani L.B., M.Pd, his second advisor, for the advice, corrections and encouragement for him until this final project was completely done.

He is also very grateful to all of his lecturers of the English Department of UNNES for all guidance and knowledge during his study at UNNES. His writer also expresses his special thanks to the three other members of the board of examiners. Dra. Malarsih, M.Sn., as the first chairman of examination, Drs. Suprapto, M.Hum., as the secretary of the examination and Amir Sisbiyanto, M.Hum, as the first examiner.

His special gratitude is forwarded to his beloved Mom and Dad for their patience and love. He also thanks to all of his family members, who gave him spirit in finishing this study. The last to people who helped me that cannot be mentioned one by one, I thank them very much. Hopefully, Allah gives his help and blessing to them.

ABSTRACT

Haris Rizqi Arifin. *2009. Analysing of The Test Items in English Final Examination for The Sixth Grade Students of Elementary Shools In South Semarang Regency In the Academic Year 2008/2009.* Final Project. English Education. English Department Languages and Arts Faculty, Semarang State University (First Advisor: Drs. Hartoyo, M.A, PhD, Second Advisor: Dr. Dwi Anggani, LB, M.Pd).

Key words: English, Achievement Test, validity, Reliability, Difficulty Level, Discriminating Power.

One way to know student's ability in using English is evaluation or test. In learning, test is a tool of evaluation which has important role to measure the teaching learning process in schools. The main purpose of this study is to analyze the English final examination items which are administered to sixth grade students of Elementray Schools in South Semarang Regency. The problem of this study is "How good are test items in final test prepared for the sixth grade students of Elementray Schools in South Semarang Regency in the academic year of 2008/2009?"

Achievement test emphasizes past progress, whereas aptitude test primarily concerns with future potentialities. Achievement test is used for assessing present knowledge and abilities. The primary goal of the achievements test is to measure past learning, that is, the accumulated knowledge and skills of an individual in a particular field.

The data used in this study were taken from the test papers and students' answer sheets. The test papers consist of 50 items in the form of multiple choices. The students' answer sheets are needed for statistical analysis to find out the quality of the items based on item analysis, validity and reliability of the test.

From the result of the analysis, the mean of validity level is 0.3250. Then, the result was consulted to the value of product moment formula at level of significance 0.05. Since the value of r circulation is more than of the table, it can be concluded that the test is valid. However, this test is reliable, with the coefficient of reliability of the whole test items is 0.946. The mean of the difficulty level is 0.83. So, the English summative test items are classified as easy items in term of their difficulty level. Then the mean of the discrimination power is 0.20, meaning that the items are still able to discriminate the clever students and the poor ones. Also the dependability is 0.963.

Based on the result, the writer suggests to the teacher as the test makers to prepare test items far in advance before they give it to the students. They should also pay attention to the writing of multiple choice items and the characteristic of a good language. Finally, the writer draws a conclusion that the items in the English final test for the sixth grade students of Elementray Schools in South Semarang Regency could still be used as an instrument of evaluation with some revisions.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In the first chapter, the writer would like to discuss background of the study, reasons for choosing the topic, statement of the problems, objective of the study, significance of the study, limitation of the study, and outline of the final project.

1.1Background of the Study

English is one the most widely used international languages in the world. Today in Indonesia, English is introduced into the curiculum and considered as the first foreign language to be taught from elementary school. English has been introduced as a local content curriculum at the elementary school. In an educational process, students or learners are expected to undergo changes. Given this view, we expect that each program, course and educational unit brings about some significant changes in the students. To find out whether the expected changes have been taken place or not it is necessary for teachers to conduct a test or an examination as one of the evaluation instrument. There are many advantages that we can acquire from the evaluation of the school program.

> Richard states that: “Evaluation, in a language teaching program, is that phase or language program development that (1) monitors the teaching process in order to ensure that system work, and (2) determines which phases of the system need adjustment when problems are expected.” (Richard, 1985:9)

Recognizing that evaluation is very important in school, teachers have to know the quality of a good test or criteria of a good test. There are some characteristics of a good test (Arikunto, 2005:53):

a. Test have high validity. A validity represents all important condition in making of test. An evaluation technique is considered has high validity if the test measures what actually to be measured.
b. Tests shoud be reliable or can be trusted. It gives consistent result, if it is tested several times. A test is reliable if the test shows constancy.
c. A test must be objective. It means that, a test has objectivity if there is no subjective factor in doing the test especially in scoring system.
d. Tests must be practical and has clear instruction.

Therefore, writer tries to analyze the item test of evaluation in final examination based on validity, reliability, discrimination power, and index difficulty because the writer hopes that he can make a good test.

1.2Reasons for Choosing the Topic

In this study, the writer would like to focus the research on the English test items used in the summative test at the elementary schools in South Semarang Regency. The reasons for choosing the topic are as follows:

a. The summative test for the sixth year students of elementary school in South Semarang Regency in academic year of 2007/2008 has never been analyzed in terms of its validity, realibility, discriminations power and difficulty level.

b. In KTSP 2006 Curriculum, the English material for each term is so substantial subject in teaching-learning process that the best items test must be selected, especially the proportion of the number of items with the material covered in each term.

c. According the recent curriculum, the evaluation of the teaching-learning process is carried out twice a year and the final examination (UAS) is held in the last year for the sixth year of elementary school students. If test constructor do not pay attention in selecting the items, the validity and the reliability of each test items will be less guaranted. For this reason, every test constructor must be careful in constructing the test items so that the result will meet the disired goal.

1.3Statement of the Problems

Through this study, the writer would like to find out the answer of the following question: "How good is the summative test made by the association of local English teachers in Semarang for the sixth year students of the elementary school in South Semarang Regency in the academic year of 2007/2008?"

A good test is a test which is arranged by considering the essential characteristics of a test. Harris (1969:13) points out "Three characteristics of a good test called: validity, reliability, and practically". In analyzing the test, the writer limits the problem further into the following questions:

a. What is the difficulty level of the test items?

b. What is the discrimination level of the test items?

c. What is the validity of the test items?

d. What is the reliability of the test items?

e. What is the dependability of the test items?

1.4Objective of the Study

The general objective of this study is to obtain an objective description of the English Final School Examination (UAS) made by the association of the local English teachers in Semarang for the sixth year students of elementary school in South Semarang Regency in academic year of 2007/2008.

The objectives are then specified into following goals:

a. To describe the value of the difficulty of the test items.

b. To describe thevalue of the discriminating power of the test items.

c. To describe the validity of the test items.

d. To describe the reliability of the test items.

e. To describe the dependability of the test items.

1.5Significance of the Study

The advantages that can be required from this study are as follows:

a. For teacher: Teacher can use the result of this study as a reference when they want to analyze test items. So, the can applied the material based on the this test items to face the next tests.

b. For test constructor: The test constructor may use it as a supplement in constructing the next tests. He or she can choose the good items to applied it in the next tests.

1.6 Limitation of the Study

There is a limitation in this final project. The writer only analyze elementary final examination. It's only analyze discrimination power, index difficulty, validity, reliability, and dependability.

1.7 Outline of the Final Project

This final project is divided into five chapters. Chapter I, the introduction, consist of general background of the study, reasons for choosing the topic, statement of the problems, objective of the study, significance of the study, limitation of the study, and the outline of the final project.

Chapter II present review of the related literature in this study. The writer is of the opinion that is important to review literature related to english testing. This chapter discussed the characteristics of a good test, item analysis, which deals with the analysis of the relevance of a progress or achievement test to the curriculum and also a brief review of the multiple choice test items.

Chapter III deals with methodology of the study, which presents the population and sample, sampling techniques, identification of the problems and techniques of data collecting.

Chapter IV presents the analysis and the discussion of research findings.

Chapter V gives the conclusion of the research and some suggestions on the basis of the researh findings.

CHAPTER II

REVIEW OF THE RELATED LITERATURE

In the second chapter, the writer would like to discuss testing, evaluation, measurement; criteria of a good test; item analysis; item difficulty, item discrimination power; validity, types of validity; reliability; dependability; types of test; achievement test, types of achievement test;multiple-choice test item.

2.1Testing, Evaluation, Measurement

A test or an examination (or "exam") is an assessment, often administered on paper or on the computer, intended to measure the test-takers' or respondents' (often a student) knowledge, skills, aptitudes, or classification in many other topics. According to Heaton (1975:1), "tests maybe constructed primarily as devices to reinforce learning and to motivate the student, or primarily as a means of assessing the student's performance as the language." Meanwhile, Valette (1977:3) argues that "testing is a topic of concern to language teachers, both those in the classroom and those engaged in administration or research."

Evaluation is one of the activities to measure and asses the level of achievement of students. Evaluation is systematic determination of merit, worth, and significance of something or someone using criteria against a set of standards. Evaluation often is used to characterize and appraise subjects of interest in a wide range of human enterprises, including the arts, criminal justice, foundations and non-profit organizations, government, health care, and other human services.

Measurement is a method to asses the student's based on the rules. Measurement is the process of obtaining the magnitude of a quantity such as length or mass relative to a unit of measurement. The term can also be used to refer to the result obtained after performing the process. Measurement is the process observing and recording the observations that are collected as part of a research effort.

Testing, Evaluation, and Measurement are three basic related concepts that we need to understand. The similarity among them is to assess the students' ability in mastering language. Test and measurement are parts of evaluation. The difference between test, evaluation and measurement can be found in the practise of asigning final marks to students at the end of a unit of work.

From the statement above, the writer can be conclude that the test is a device to asses the student's ability in teaching learning process. Through a test, teachers can get information about students achievement. Evaluation is one of the activities to measure and asses the level of achievement of students. It is important to have a good evaluation. Measurement is a method to asses the student's based on the rules.

2.2Criteria of a Good Test

"A test has important role in the teaching and learning process as an integral part of the instructional program that provides information that serves as a basis for a variety of educational decisions" (Fahmalatif, 2002:9). As stated by Madsen (1983:3), "testing is an important part of every teaching and learning experience."

Based on Brown (2004:19-30), "there are five criteria for testing a test: practically, reliability, validity, authenticity and washback." Here, the focus is on validity and reliability because the validity and reliability level is very significant.

2.3Item analysis

The analysis of students response to objective test item is powerful tool for test improvement. Reexamining each test item to discover its strength is known as item analysis. Item analysis begins after the test has been scored. According to Ebel (1991:225) a classroom teacher who chooses to complete the procedures by hand would follow these six steps:

a. Arrange the scored test papers or answers sheets in order from highest to lowest.
b. Identify an upper and a lower group separately. The upper group is the highest scoring 27 percent (one-fourth) of the group and the lower group is an equal number of the lowest scoring of the total group.
c. For each item, count the number of examinees in the upper group that choose each response alternative. Do a separate, similar tally for the lower group.
d. Record these counts on a copy of the test at the end of the corresponding response alternatives. (The use of colored pencils is recomended.)

e. Add the two counts for the keyed response and divide this sum by the total number of students in the upper and lower groups. Multiply this decimal value by 100 to form percentage. The result is an estimate of the index of the index of item difficulty.
f. Substract the lower group count from the upper group count for the keyed response. Divide this difference by the number of examinees in one of the groups (either group since both are the same size). The result, expressed as a decimal, is the index of discrimination.

Item analysis usually concentrates on three vital features: level of difficulty, discriminate and the effectiveness of each alternative. Item analysis somtimes suggests why an item has not functioned effectively and how it might be improved.

2.3.1 Item difficulty

Item difficulty is indicated by the percentage of the students who got the item correct. The more difficult of item is the fewer will be the students who select the correct the opinion. And the easier the test is the more will be the students who selected the correct one.

> Mehrens and Lehmann (1984:81) say: "The concept of difficulty or the decision the test should depends upon a variety of factors such as (1) the purpose of the test, (2) the ability level of students, and (3) the age or grade of the students."

2.3.2 Item Discrimination Power

The discriminating power of a test is its ability to differentiate between students who have achived well (the upper group) and those who have achieved poorly (the lower group). To estimate item discriminating power is by comparing the number of students in the upper and lower group who answered the item correctly.

> According to Gronlund (1982:103) "the computation of item discriminating index (D) for each item can be done by substracting the number of students in the lower group who get the item right (L) from the number of students in the upper group who get the item right (U) and divided by one half on the total number of students included in the item analysis (1/2 T)."
>
> Tinambuan (1988;145) says that "the discrimination index can take values from 0.00 – +1.00. the higher the D value for an item, the better that item discriminated. Any item which has a D value of +0.40 or above is considered to be good in discriminating student differences. D values between +0.20 and +0.39 are usually considerd to be satisfactory, but items with the lower values in this range should be reviewed and revised to make them more effective discriminators."

2.4Validity

Validity is the most important variable of a measurement instrument. Brown (2004:22) states that "Validity is the most complex criterion of an effective test and arguably the most important principle."

2.4.1 Types of Validity

Basically, there are many types of validity according some experts. According to Brown (2004:22-30), validity is divided into five types of evidence:

a. Content-Related Evidence
   If a test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behavior that is being measured, it can claim content-related validity,

often popularly referred to as content validity (Mousavi:2002, Hughes:2003 quoted by Brown, 2004:22).

b. Criterion-Related Evidence
   A second form of evidence of the validity of a test may be found in what is called criterion-related evidence. Also referred to as criterion-related validity, or extent to which the "criterion" of the test has actually been reached.
c. Construct-Related Evidence
   Construct-Related Evidence commonly referred to as construct validity. A construct is an y theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Construct validity is a major issue in validating large-scale standardized test of profeciency.
d. Consequential validity
   Consequential validity encompasses all the consequences of the test, including such consideration as its accuracy in measuring intended criteria, its impact on the preparation of test-takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use.
e. Face Validity
   Gronlund (1982:210) quoted by Brown (2004:26) says an important faced of consequential validity is the extent to which "students view the assessment as fair, relevant, and useful for improving learning," or what is popularly known as face validity. "Face validity refers to the degree to which a test looks rights, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgement of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers" (Brown adapted from Mousavi, 2002:244)

From the explanations above,the writer concludes that validity is one of the most important criteria of a good test. There are types of validity. They are content-related validity, criterion-related validity, construct related-evidence, consequential validity, and face validity.

2.5Reliability

Reliability is the extent to which result can be considered consistent or stable, that is how consistent test scores or other evaluation results are from mesurement to another.

There are three ways of estimating this method. First is odd and even method: a method for estimating reliability of a test by giving a single administration of one form of the test then dividing the items into halves usually by separating odd and even number items. The second method is obtaining two scores for each individual. Then the reliability coefficient can be determined by computing the correlation between them. The third method is called Kuder-Richardson Method. This method measures the extend to which items within one form of the test have as much in common with one another as do the items in that one form with corresponding items in an equivalent form.

Reliability is the extent to which the result can be considered consistent. A test should be reliable because unreliable test might produce different scores if it is taken again.

2.6Dependability

In Indonesia we are using Criterion Reference Test (CRT) to make a test, not using Norm Reference Test (NRT).

> Based on Brown (2005: 199), “the terms agreement and dependability are used exclusively for estimates of the consistency of CRT’s, while the

> term reliability is reserved for NRT consistency estimates. This distinction helps teachers and testers keep the notions of NRT reliability separate from the ideas of CRT agreement and dependability.”

2.7Type of Test

Tests motivete and direct student learning because tests guide student learning and help determine how students will prepare for a test.according to Vallete (1977:5-6) there are four types of test. They are:

a. The aptitude test
The aptitude test is conceived as a prognostic measure that indicates whether a student is likely to learn a second language readily.
b. The progress test
The progress test measures how much the student has learned in a specific course of instruction.
c. The achievement test
The achievement test is similar to the progress test in that it measures how much the student has learned in the course of second language instruction.
d. The proficiency
The proficiency test also measures what students have learned, but the aim of the proficiency test is to determine wheter this language ability corresponds to specific language requirements.

2.8Achievement Test

Achievement test plays an important role in all types of instructional programs. It is the most widely used method of assessing students’ achievement in classroom instruction and it is indispensable procedure in individualized and program instruction.

Achievement test is used to assessing presents knowledge and abilities. The primary goal of the achievement test is to measure past learning, that is the accumulated knowledge and skills of an individual in a particular field or fields. Brown (2004:47) states that "an achievement test is related directly to classroom lessons, units, or even a total curriculum."

2.8.1 Types of Achievement Test

Tinambuan (1988:7-9) says, there are four types of achievement test which are very commonly used by teachers in the classroom:

a. Placement test
   Placement test is designed to determine the pupil performance at the beginning of instruction.
b. Formative test
   Formative test is intended to monitor learning progress during the instruction and to provide continous feedback to both pupil and teacher concerning learning successes and failures.
c. Diagnostic test
   Diagnostic test is intended to diagnose learning difficulties during instruction.
d. Summative test
   The summative test is intended to show the standard which the students have now reached in relation to other students at the same stage.

Based on the statements, the writer defines that achievement test is used for assessing presenst knowledge and abilities. There are some types of achievement test, they are: placement test, formative test, diagnostic test, and summative test.

2.9Multiple-choice Test Item

Heaton (1975:14) says that "multiple-choice is now widely regarded as being one of the most useful of all objectives item types." Although it is among the most difficult of all objective item types to construct, it is simple to score and administer. Valette (1977:7) states that "multiple choice test items are designed to elicit specific responses from the students."

The multiple-choice item consists of two parts. They are: stem or lead, which is either a direct question or incomplete statement. The students will have to answer or complete one alternative. Alternatives may consists of two or more coices or responses of which one is the answer and the others are distracters, that are, the incorect responses. The function of distractors is to distract those students who are uncertain of the answer.

An attractive feature of multiple-choice questions is that they are particularly easy to score. Multiple-choice tests are also valuable when the test sponsor desires to have immediate score reporting available to the examinee; it is impossible to provide a score at the end of the test if the items are not actually scored until several weeks later. This format is not, however, appropriate for assessing all types of skills and abilities.

CHAPTER III

METHOD OF INVESTIGATION

In the third chapter, the writer would like to discuss population and samples, sampling technique identification of the problems, techniques of data collection, and technique of data analysis.

3.1 Population and Sample

Margono (2003:18) said that “a population is defined as a complete set of individuals or subject having coming observable characteristics.” “The population is the establishment of boundary condition that specify who shall be included in or excluded from the population” (Tuckman, 1978:117).

The population of this study was the sixth grade students of Elementary School in South Semarang Regency in the academic year of 2007/2008. The writer thought that the number of Elementary School in South Semarang Regency is too big for this purpose, so the writer took only five schools as the sample.

3.2 Sampling Technique

In order to make this study effective, researcher had to select sample. Sample is part of population, which represent the population.

According to Brink (1974), “random sampling refers to the process of drawing a random sample of individuals of some population.” In this study, the writer used random sampling technique to take samples. In the random sampling technique, each number has an equal chance of being selected for the sample. The writer took twenty students of each school to be taken as samples. So there were a hundred students taken from five different schools.

In selecting twenty students of each school, the writer took the procedure called lottery method. This method is an objective selection, he did it by writing down the order number of the student’s names list on a small piece of paper, and then the piece of papers was rolled and let the ten rolls of the paper drop out of the glass one after another. Although the steps in taking samples are very simple, many researchers have to adapt the random sampling techniques as one way to select the samples since it is not influenced by thought and feelings.

3.3 Identifications of the Problems

The fact of the analysis results shows that the most of Elementary School teachers do not know how to construct a good test.

Based on the fact above, there are four problems related to the teacher-made English test items. The problems are the difficulty level is not suitable, in this case,many questions are too easy. Many questions have low discrimination power. The validity and the reliability are low or too high.

3.4 Technique of Collecting Data

The technique of data collection in this study involves several steps, those are:

In this study the intended test is the final examination English summative test for students of Elementary School in South Semarang Regency in the academic year of 2007/2008. This test was held on May 2008. The data here are the form of students' answer sheets.

The writer selected fifteen elementary Schools in South Semarang Regency to get the required data. These schools are located near the writer house, so these schools can be reached easily.

Before the test was administrated to the students, the writer had contracted the English teacher of the selected school to ask the students' answer sheets to assume that they were not use anymore. Then, he began to analyze them.

3.5 Technique of Data Analysis

The data to be analyzed in this study were taken from the students' answer sheets of the final examination of the English summative test for the sixth grade of elementary school in South Semarang Regency in the academic year of 2007/2008. They were used to analyze the quality of the test items.

The purpose of this item analysis is to identify the quality of each item, whether they belongs good items, moderate items, or bad items. Through the item analysis, we an also find information about the weakness or the shortcoming of the items. Here, the item analysis consists as the following:

3.5.1 Difficulty Level Analysis

A good test item is an item which is not too difficult or too easy. The difficulty of the test items is the percentage of students who get the right items. Here the index of items difficulty level (P) used Nitko formula to analyze. Then, the writer divided the level of items difficulty (P) into three categories. The criteria of item difficulty level could be seen in the table below:

No	**Index Difficulty Level**	**The categories**
1	0.00-0.30	Difficult
2	0.30-0.70	Moderate
3	0.70-1.00	Easy

(Heaton, 1975:172)

The formula is:

$$P = \frac{R}{T}$$

Where: P= difficulty level or index of difficulty.

R= the number of students who respond correctly to an item.

T= the total number of students who respond to the item.

(Nitko, 1983:228)

3.5.2Discrimination Power Analysis

The discrimination power of the test items tell how well the item performs in separating the upper group and the lower group. The formula to compute item discrimination is as follows:

$$D = \frac{RU - RL}{1/2T}$$

Where: D = the index of discrimination power

RU = the number of students in the upper group who answer the items correctly.

RL = the number of students in the lower group who answer the items correctly

½ T = one half of the total number of students included in the items analysis

The criteria of item discrimination power could be seen in the table below which is proposed by Ebel and Frisbie (1991: 232)

Discrimination index	**Item evaluation**
0.70≤ DP ≤ 1.00	Excellent
0.40< DP ≤ 0.70	Good
0.20< DP ≤ 0.40	Satisfactory
0.00< DP ≤ 0.20	Poor

3.5.3Analysis of Validity

Validity refers to whether or not a test measures what it should be measure. In this study the writer used criterion-related validity (validity coefficient to determine whether the test items were valid or not). To find the coefficient validity, he used Peargon Product moment formula. The formula is as follows:

$$rx = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\{N\sum x^2 - (\sum x)^2\}\{N\sum y^2 - (\sum y)^2\}}}$$

Where rxy : correlation index

x : the total score

N : the total number of the respondent

Σ : the sum

(Hinkle, Durs, Wiersma, 1979:98)

There are two ways to determine the value of correlation coefficient. First, by interpreting the value of calculation with the following criteria:

$0.810 \leq rxy \leq$=very high validity

$0.610 \leq rxy \leq$=high validity

$0.410 \leq rxy \leq$=moderate validity

$0.210 \leq rxy \leq$=low validity

$0.000 \leq rxy \leq$=very low validity

(Bloom, 1981:152)

According to Arikunto (1992:161) if the value of calculation is lower than the critical value on the table, so the correlation is not significant or we can say that the items is not valid and so vice versa.

3.5.4Analysis of Reliability

In this study, the writer used the Kuder Richardson formula in estimating the reliability of the formula. The formula is:

$$r = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{s^2}\right)$$

Where r: reliability coefficient of the test items

k: number of item in the test

p: the difficulty index

q: the portion of the students give the wrong answer (q=1-p)

s: the variance of the total test scores

(Phopam, 1981:1430)

The formula to calculate the variant is:

$$s^2 = \frac{\Sigma Y^2}{N}$$

Where s : the variance

Σ : the sum of

Y : the total score

N : the total respondent

The result of the reliability calculation is constructer to the value of critical production product moment on the table: we can say that the item is not reliable. On the other hand, the item is reliable if the value of calculation is more than the value of the table.

3.5.5Analysis of Dependability

In this study, the writer used the following computation in estimating dependability of the instrument. The formula is:

$$\varphi = \frac{\frac{n\mathrm{Sp}^2}{n-1}[K - R20]}{\frac{n\mathrm{Sp}^2}{n-1}[K - R20] + \frac{\mathrm{Mp}(1 - \mathrm{Mp}) - S_p^2}{k-1}}$$

(Brown, 2005:208)

Where:

n : number of persons who took the test

k : number of items

M_p : mean proportion scores

S_p : standard deviation of proportion scores

K–R20: Kuder-Richardson formula 20 reliability estimate

CHAPTER IV

ANALYSIS AND DISCUSSION

In the fourth chapter, the writer would like to discuss the result of analysis, analysis of validity, analysis of reliability, analysis of difficulty level, analysis of discrimination power and discussion.

4.1 Analysis

The goal of this study is to analyze items of the English final examination for the sixth grade Elementary Students in South Semarang Regency in the academic year of 2007/2008. The analysis consists of four aspects, namely the difficulty level, discrimination power, validity, reliability,and dependability of the test.

Item analysis aims to identify good, moderate, and poor items. Though item analysis, we get information about the shortcomings of the items and how to revise them. From data analysis of the English final examination for the sixth grade students of Elementary Schools in South Semarang Regency in academic year 2007/2008, the writer obtained the following results.

4.1.1 Difficulty Level

The difficulty of the test items is the percentage of students who got the right items. Here the index of item difficulty level (P) used Nitko formula to analyze. Then the writer divided the level of items difficulty (P) into three categories. The criteria of item difficulty level could be seen in the table below:

No	Index Difficulty Level	The categories
1.	0.00 – 0.30	Difficult
2.	0.31 – 0.70	Moderate
3.	0.71 – 1.00	Easy

From the table in appendix 4, the result of the data analysis of the item Difficulty Level shows as follows:

No.	Criteria	Items Number	Total	Percentage
1.	Difficult Items		-	-
2.	Moderate Items	2, 3, 5, 6, 8, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20, 22, 23, 26, 27, 28, 30, 31, 34, 36, 39, 40, 42, 43, 44, 45, 46, 47, 48, and 49.	34	68%
3.	Easy Items	1, 4, 7, 12, 14, 21, 24, 25, 29, 32, 33, 35, 37, 38, 41, and 50.	16	32%

From the test, the mean of their difficulty level is 0.83. So, the English final examination items are classified as Easy items in terms of their difficulty level. Items, which are considered very easy, can still be used in a test to encourage and motivate the poor students. The example of computation of item difficulty level is listed in apendix 4.

4.1.2 Discrimination Power

The discrimination power of the test items tells how well the item performs in separating the upper group and the lower group. The writer divided the discrimination power of the test items into four categories. The criteria are as follows:

Discrimination index	**Item evaluation**
0.70≤ DP ≤ 1.00	Excellent
0.40< DP ≤ 0.70	Good
0.20< DP ≤ 0.40	Satisfactory
0.00< DP ≤ 0.20	Poor

From the table in appendix 1, the result of data analysis can be seen on this table.

No.	Criteria	Items Number	Total	Percentage
1.	Excellent Items		-	-
2.	Good Items	2, 3, 5, 6, 8, 9, 16, 17, 19, 23, 27, 28, 30, 31, 34, 36, 37, 39, 42, 44, 47, and 49.	22	44%
3.	Satisfactory Items	4, 7, 10, 11, 12, 13, 14, 15, 18, 20, 21, 22, 24, 25, 26, 29, 32, 33, 35, 38, 40, 41, 43, 45, 46, 48, and 50	27	55%
4.	Poor Items	1	1	1%

4.1.3 Validity

$$r_{xy} = \frac{(100 \text{ x } 2912) - (83 \text{ x } 3330)}{\sqrt{\{(100 \text{ x } 83)(83)^2\}\{(100 \text{ x } 125608)(3330)^2]}}$$

r_{xy} = 0.3250

On a = 5% with N= 100 it is obtained = 0.195

The Pearson's product moment formula is used to calculate the validity level of the test items since the value of r calculation is more than the r table ($r_c > r_t$), the item is valid and vice versa. For N = 100 with the significance level 0.05, the value of r on the table is 0.195 (see appendix 2). From thevalidity calculation, the writer got the results as follows:

No	Criteria	Number	Percentage
1.	Valid Items	50	100%
2.	Invalid Items	-	-

There are all of test items, which fulfill the requirements of validity. So, there is no test item which do not fullfill the requirements of the validity.

From the table, we fond that the validity of the items is 0.3250. The example of the computation of item validity is listed in apendix 2.

4.1.4 Reliability

$$r = \left[\frac{50}{50-1}\right]\left[\frac{147,190 - 10,7455}{147.1900}\right]$$

$$r = 0.946$$

As the writer has stated in the previous chapter, the coefficient of reliability of test items is found by applying the Kuder-Richardson 21 formula. From the computation, it is found that the coefficient of the test is 0.949. The result is then consulted to the table of r product moment values at level of significance of 0.05.

It is found that the value of r 0.195 for N 100. Since value of r calculation is more than that of table (r), it can be conclude that the test items used in English final examination for sixth grade students in Elementary Schools in South Semarang Regency in academic year 2007/2008 is reliable. The computation of the reliability coefficient is listed in apendix 3.

4.1.5 Dependability

$$\varphi = \frac{\frac{100 \times (0.91)^2}{100-1}[0.946]}{\frac{100 \times (0.91)^2}{100-1}[0.946] + \frac{0.746(1-0.746)-(0.91)^2}{50-1}}$$

$$\varphi = 0.963$$

What is necessary for calculating this coefficient of dependability is the number of students, number of items, mean of the proportion scores, standard deviation of the proportion scores, and the K-R20 reliability estimate. The result of 0.963 means that the scores on the test about 96 percent dependable for testing this particular domain. This fact has one important implication: because K-R21 (0.946) is lower than φ (0.963), then K-R21 can serve as a conservative “rough and ready” underestimate of the domain-referenced dependability (φ) of a test.

4.2 Discussion

The goal of the writing of this final project is to identify the quality of each item, whether it can be classified as good, moderate, or poor item. It later can be determined which items can still be used, can be used with revision, or should be dropped. From the point of view of difficulty level, a good item is an item, which is not too easy or not too difficult. From the discrimination power of view, a good item is an item that can be discriminate between students from the upper group and the students from the lower group.

Based on the result of item analysis which includes the analysis of difficulty level, discrimination power, validity, and reliability of the items (see appendix 1), this test items can be used in English final examination with several revison.

CHAPTER V

CONCLUSION AND SUGGESTION

In the fourth chapter, the writer would like to discuss the result of analysis, analysis of validity, analysis of reliability, analysis of difficulty level, analysis of discrimination power and discussion.

5.1Conclusion

According to the result of the analysis of the fifty test items administrated to five schools in South Semarang Regency had helped the writer to come the following conclusions.

1) The mean of the index of difficulty for the examined was 0.83. It means that the P value is between 0.71 – 1.00, which put the test in the easy position. As a result, on the whole the English Final examination have met requirements of an easy test in terms of the difficulty level.
2) In analysis of the item discrimination power, it was found out that the mean of the D value was 0.40, which put the English Final test items in the satisfactory test.
3) In analysis of item validity, it was found that the value of validity of the whole test items is 0.3250. It means that the test is valid.

4) By applying KR-21 formula, the writer found that the coefficient of reliability of the whole test item is 0.946. It means that the test as a whole had high reliability, and we can use the test items as the instrument of evaluation again if we want to.
5) In analysis of dependability, it was found that the value of dependability of this domain is 0.963.
6) Finally, the writer draws a conclusion that the items in the English Final Examinatin for Sixth grade students of Elementary School in South Semarang Regency in academic year 2007/2008 could still be used as an instrument of evaluation with some revisions.

5.2Suggestions

Constructing good language test items is not an easy task. Based on the conclusions above, the writer would like to offer the following suggestions.

First, the test constructors should know about the characteristics of good language test, especially procedure of determining difficultylevels and discrimination power.

Second, items that still can be used should be revised and save.

Third, items which have negative value should be discarded, it means because of performing of the lower group is better than the upper group.

There are some points to be considered in constructing test items.

(1) Prepare the test item far away, before they are administrated to the students, it will helps the constructor develop good test item.

(2) Write each test item related to the intended learning outcomes to be measured.

(3) The difficulty level of each item should be match with the students ability

And finally, the writer suggestion this test should be used in the English Final Examination. It can be used unless it has makes some revisions and the writer hopes that the result on this item analysis could be used as an example in analyzing other test items, and encourages other teachers to do research on the same object.

Bibliography

Arikunto, S. 2005. *Dasar – Dasar Evaluasi pendidikan* (Revised Ed.). Jakarta: Bumi Aksara.

Bloom, Benjamin., Krathwol, David R., and Marsia, Bertram B. 1964. *Taxonomy of Educational., Objectives Book 2: Taxonomy Domain.* London: Longman, Green and Co. Ltd.

Brown, H.D. 2004. *Language Assessment: Principles and Classroom Practices*. New York: Longman.

Brown, H.D. 2004. *Testing in Language Programs: Principles and Classroom Practices*. New York: McGrawHill Inc.

Bygate, M. 2000. *Researching Pedagogic Tasks: Second Language Learning, teaching and Testing.* Horlow: Pearson Education Limited.

Ebel, R.L and D.A Frisbie. 1979. *Essential for Educational Measurement.* New Jersey: Prenctice – Hall: University of Lawn.

Ebel, R.L and D.A Frisbie. 1991. *Essential for Educational Evaluation*. Philipines: Addison – Wersley Publishing Company.

Fahmalatif, Farida.2002. *An Analisys of English Summative Test for First Year SMU Students in the First Term of Academic 2001/2002.* Unpublished. Semarang: UNNES.

Gay. L. R. 2008. *Educational Research Computation for Analysis and Application*. Columbus: Charles Merril Publishing.

Gronlund, Norman E. 1976. *Measurement and Evaluation In Teaching*. New York: Macmillan Publishing Co., Inc, University of Illnois.

Gronlund, Norman E. 1982. *Constructing Achievement Test*. New Jersey: Prentice – Hall. Inc. University of Illnois.

Harris, D.P. 1969. *Testing as a Second Language*. New York: Mc Graw – Hill.

Hartoyo. 2008. (*Unpublished Hand Out) Introduction to Educational Research.* Semarang: UNNES.

Heaton, J.B. 1975. *Writing English Language Test*. England: Longman Group Limited.

Hinkle, D.E., Jurs, S.G, Wiersma, W. 1979. *Applied Statistics for the Behavioral Sciences*. Hough Company.

Lado, Robert. 1961. *Language Testing*. New York: Oxford University Press.

Madsen, Horald S. 1983. *Technique in Testing*. Hongkong: Oxford University.

Margono, Drs. 2003. *Metodologi Penelitian Pendidikan*. Jakarta: Rineka Cipta.

Nitko, A.J. 1983. *Educational Tests and Measurement an Introduction*. London: Harcourt Brace Jovanovich Inc.

Oller, J. W. 1979. *Language Test at School*. London: Longman Group Limited.

Popham, James W. 1981. *Modern Educational Measurement.* London: Prentice Hall Inc. Englewood Cliffs.

Richard, C. A. 1973. *Educational psychology: The Science of Instruction and Learning.* USA: Happer and Raw, Publisher, Inc.

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

Tuckman, B.W. 1978. *Consducting Educational Research*. London: Harcout Brace Jacobovitz.

Vallete, R.M. 1977. *Modern Language Testing*. New York: Harcout Brace Jacobovitz

The Computation of Item Validity Test

Formula

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2 - (\Sigma X)^2\}\{N\Sigma Y^2 - (\Sigma Y)^2\}}}$$

The item test is valid if $r_{xy} > r_{tabel}$

The following is the example of counting the validity of item number 1, and for the other items will use the same formula.

No.	Code	X	Y	X^2	Y^2	XY
1	T-01	1	50	1	2500	50
2	T-02	1	50	1	2500	50
3	T-03	1	50	1	2500	50
4	T-04	1	50	1	2500	50
5	T-05	1	50	1	2500	50
6	T-06	1	50	1	2500	50
7	T-07	1	50	1	2500	50
8	T-08	1	50	1	2500	50
9	T-09	1	50	1	2500	50
10	T-10	1	50	1	2500	50
11	T-11	1	48	1	2304	48
12	T-12	1	48	1	2304	48
13	T-13	1	48	1	2304	48
14	T-14	1	48	1	2304	48
15	T-15	1	47	1	2209	47
16	T-16	1	47	1	2209	47
17	T-17	1	47	1	2209	47
18	T-18	1	47	1	2209	47
19	T-19	1	47	1	2209	47
20	T-20	1	47	1	2209	47
21	T-21	1	45	1	2025	45
22	T-22	1	45	1	2025	45
23	T-23	1	45	1	2025	45
24	T-24	1	45	1	2025	45
25	T-25	1	45	1	2025	45
26	T-26	1	43	1	1849	43
27	T-27	1	43	1	1849	43
28	T-28	1	42	1	1764	42
29	T-29	1	42	1	1764	42
30	T-30	1	42	1	1764	42
31	T-31	0	42	0	1764	0
32	T-32	1	42	1	1764	42
33	T-33	1	42	1	1764	42
34	T-34	0	42	0	1764	0
35	T-35	1	40	1	1600	40

No.	Code	X	Y	X^2	Y^2	XY
36	T-36	1	40	1	1600	40
37	T-37	1	40	1	1600	40
38	T-38	1	40	1	1600	40
39	T-39	1	40	1	1600	40
40	T-40	1	40	1	1600	40
41	T-41	1	40	1	1600	40
42	T-42	1	38	1	1444	38
43	T-43	0	38	0	1444	0
44	T-44	1	38	1	1444	38
45	T-45	1	38	1	1444	38
46	T-46	1	38	1	1444	38
47	T-47	0	38	0	1444	0
48	T-48	1	34	1	1156	34
49	T-49	1	34	1	1156	34
50	T-50	1	34	1	1156	34
51	T-51	0	34	0	1156	0
52	T-52	0	34	0	1156	0
53	T-53	1	34	1	1156	34
54	T-54	0	34	0	1156	0
55	T-55	1	30	1	900	30
56	T-56	1	30	1	900	30
57	T-57	1	30	1	900	30
58	T-58	1	30	1	900	30
59	T-59	1	30	1	900	30
60	T-60	1	30	1	900	30
61	T-61	1	30	1	900	30
62	T-62	1	30	1	900	30
63	T-63	1	30	1	900	30
64	T-64	1	30	1	900	30
65	T-65	1	30	1	900	30
66	T-66	1	30	1	900	30
67	T-67	1	30	1	900	30
68	T-68	1	30	1	900	30
69	T-69	1	30	1	900	30
70	T-70	1	30	1	900	30
71	T-71	1	25	1	625	25
72	T-72	1	25	1	625	25
73	T-73	1	25	1	625	25
74	T-74	1	25	1	625	25
75	T-75	1	20	1	400	20
76	T-76	1	20	1	400	20
77	T-77	1	20	1	400	20
78	T-78	1	20	1	400	20
79	T-79	1	20	1	400	20
80	T-80	1	20	1	400	20
81	T-81	0	18	0	324	0
82	T-82	1	18	1	324	18

No.	Code	X	Y	X^2	Y^2	XY
83	T-83	0	18	0	324	0
84	T-84	1	18	1	324	18
85	T-85	0	16	0	256	0
86	T-86	1	16	1	256	16
87	T-87	0	16	0	256	0
88	T-88	0	15	0	225	0
89	T-89	1	15	1	225	15
90	T-90	0	15	0	225	0
91	T-92	0	15	0	225	0
92	T-93	1	15	1	225	15
93	T-94	0	15	0	225	0
94	T-91	1	14	1	196	14
95	T-95	1	14	1	196	14
96	T-96	0	14	0	196	0
97	T-97	1	14	1	196	14
98	T-98	0	14	0	196	0
99	T-99	1	14	1	196	14
100	T-100	1	11	1	121	11
Σ	□	83	3330	83	125608	2912

By using that formula, we obtain that :
r_{xy}=0,3250

On a = 5% with N= 100 it is obtained = 0,195
Because of $r_{xy} > r_{tabel}$, so the item number 1 is Valid.