

**THE VALIDITY, RELIABILITY, LEVEL OF DIFFICULTY
AND APPROPRIATENESS OF CURRICULUM OF
THE ENGLISH TEST**

**(A Comparative Study of The Quality of English Final Test of The First
Semester Students Grade V Made By English KKG of Ministry of Education
and Culture and Ministry of Religion Semarang)**



**A THESIS
In Partial Fulfillment of the Requirements
for Master's Degree in Linguistics**

**Athiyah Salwa
NIM. 13020210400002**

**POSTGRADUATE PROGRAM OF LINGUISTICS
DIPONEGORO UNIVERSITY
SEMARANG**

2012

A Thesis

**THE VALIDITY, RELIABILITY, LEVEL OF DIFFICULTY AND
APPROPRIATENESS OF CURRICULUM OF
THE ENGLISH TEST**

**(A Comparative Study of The Quality of English Final Test of The First
Semester Students Grade V Made By English KKG of Ministry of Education
and Culture and Ministry of Religion Semarang)**

Submitted by:

Athiyah Salwa

NIM. 13020210400002

Approved by:

Advisor,



Dr. Suwandi, M.Pd

NIP. 19520815 198303 1 003

Master's Program of Linguistics

Head,



J. Herudjati Purwoko, Ph.D

NIP. 19530327 198103 1 006

**THE VALIDITY, RELIABILITY, LEVEL OF DIFFICULTY AND
APPROPRIATENESS OF CURRICULUM OF
THE ENGLISH TEST**

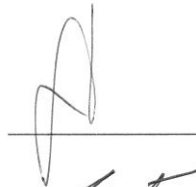
**(A Comparative Study of The Quality of English Final Test of The First
Semester Students Grade V Made By English KKG of Ministry of Education
and Culture and Ministry of Religion Semarang)**

Submitted by:
Athiyah Salwa
NIM. 13020210400002

VALIDATION

Approved by
Srata II Thesis Examination Committee
Master's Degree in Linguistics
Postgraduate Program Diponegoro University
On Thursday, August 16th 2012

Chairman
Dr. Suwandi, M.Pd.



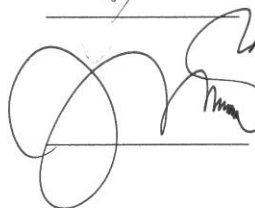
First Member
J. Herudjati Purwoko, Ph.D



Second Member
Dr. Nurhayati, M.Hum



Third member
Dr. Deli Nirmala, M.Hum



ACKNOWLEDGEMENT

Alhamdulillah, praise to the Almighty Lord the Greatest God, Allah SWT who gives the mercy, bless, and the gift and shows me the right way and inspiration during the writing process of this thesis. Shalawat and salutation go to Rasulullah SAW who is always admired by his followers including me. On this occasion, the writer would like to thank all those people who have contributed to the completion of this study report.

The deepest gratitude and appreciation are extended to Dr. Suwandi, M.Pd as the writer's advisor. He continually and convincingly conveyed a spirit of adventure in regard to research. Without his guidance, suggestion and persistent help this final project would not have been possible to complete.

My deepest gratitude is also addressed to my beloved parents who always give their never ending love, caring, support and blessing in my life. They encourage and inspire me to what I have to be.

The writer's deepest thank also goes to the following:

1. J. Herudjati Purwoko, Ph.D., the Head of Master's Program in Linguistics
Diponegoro University
2. Dr. Nurhayati, M.Hum, and Dr. Deli Nirmala, M.Hum, the thesis examiners.
3. The staff of Magister Linguistics of Diponegoro University.
4. Zaumi Ahmad, S.Pd.I as Principal of MI Darus Sa'adah Semarang and Dyah Kurniastity, S.Pd as principal of SDIT Al Kamilah who gave me permission to try out the test in their institution.

5. To all of my family, teachers and staffs at Darus Sa'adah institution
6. To the one I called Ta Ta who has given his support and ideas in a number of ways.
7. Arum Budiarti, my best friend who never says tired of supporting and accompanying me during our study.
8. All friends in Magister Linguistics Program of Diponegoro University, academic year of 2010/2011 and 2011/2012

The writer realizes that this thesis is still far from being perfect. She therefore will be glad to receive any constructive criticism and recommendation to make this thesis better.

Finally, the writer expects that this thesis will be useful to the reader who wishes to learn something about designing a good test.

Semarang, August 2012

The writer

CERTIFICATION OF ORIGINALITY

I hereby declare that this submission is my own work and that, to the best of my knowledge, and belief. This study contains no material previously published or written by another person or material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institutes of higher learning, except where due acknowledgment is made in the text of the thesis.

Semarang, 16 August 2012

Athiyah Salwa

TABLE OF CONTENTS

	Page
COVER	i
APPROVAL	ii
VALIDATION	iii
ACKNOWLEDGMENT	iv
CERTIFICATION OF ORIGINALITY	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF DIAGRAMS	xii
LIST OF APPENDICES	xiii
ABSTRACT	xv
CHAPTER I INTRODUCTION	1
1.1 Background of the Study	1
1.2 Identification of the Problems	5
1.3 Statements of the Problems	6
1.4 Objectives of the Study	7
1.5 Significances of the Study	8
1.6 Scope of the Study	8
1.7 Definition of the Key Terms.	9
1.8 Research Hypothesis	10
1.9 Outlines of the Study Report	10

CHAPTER II REVIEW OF THE RELATED LITERATURE 12

2.1 Previous Studies	12
2.2 School-Based Curriculum (KTSP)	14
2.3 Teachers Work Group (KKG)	19
2.4 Language Testing and Assessment	21
2.5 Types of Assessment and Testing.....	22
2.5.1 Multiple-Choice Test.	24
2.5.2 Short Answer Items.....	25
2.5.3 Essay Test-Items.....	27
2.6 Characteristics of a Good Test	29
2.7 Item Analysis	30
2.7.1 Validity.....	31
2.7.2 Reliability.....	32
2.7.3 Level of Difficulty.....	32
2.7.4 Discriminating Power.....	34
2.7.5 Answer of Question Form.....	36

CHAPTER III RESEARCH METHODOLO 38

3.1 Research Design	38
3.2 Population and Sample... ..	39
3.3 Research Instruments	39
3.3.1 Curriculum Checklist	39
3.3.2 Characteristics of a Good Test Checklist	40
3.3.3 Paper Test Question	40

3.3.4 Students' Answer Sheet	40
3.3.5 Data and Source of the Data	41
3.4 Method of Collecting Data.....	41
3.5 Method of Analyzing Data	42
3.5.1 Quantitative Data Analysis	42
3.5.2 Qualitative Data Analysis	48
CHAPTER IV FINDINGS AND DISCUSSIONS	51
4.1 The Findings of Quantitative Aspect	51
4.2 The Findings of Qualitative Aspect	53
4.3 Quantitative Analysis	55
4.3.1 Analyzing Validity	55
4.3.2 Analyzing Reliability	56
4.3.3 Analyzing Index of Difficulty	57
4.3.4 Analyzing Index of Discrimination	60
4.3.5 Analyzing the Distractors' Distribution	64
4.4 Qualitative Analysis	67
4.4.1 Analyzing Face Validity	67
4.4.2 Analyzing the Appropriateness of Curriculum	68
4.4.3 Analyzing the Characteristics of a Good Test and Language Use.....	79
CHAPTER V CONCLUSION AND SUGGESTION	89
5.1 Conclusion	89
5.2 Suggestions	91

BIBLIOGRAPHY	93
---------------------------	-----------

APPENDICES	95
-------------------------	-----------

LIST OF TABLES

- Table 2.1 : Competence Standard and Basic Competence of Fifth Grade of Elementary School
- Table 3.1 : Curriculum Checklist
- Table 3.2 : Checklist of the Observation of the Characteristics of a Good Test
- Table 4.1 : The Result of the Quantitative Analysis of Test-pack 1
- Table 4.2 : The Result of the Quantitative Analysis of Test-pack 2
- Table 4.3 : The Result of the Analysis of the Appropriateness of Test-Items to Curriculum

LIST OF DIAGRAMS

Diagram 2.1 : Interval Scale for Essay-test items Scoring

Diagram 4.1 : Difficulty Index of Test-Pack 1

Diagram 4.2 : Difficulty Index of Test-Pack 2

Diagram 4.3 : Discrimination Power Index of Test-Pack 1

Diagram 4.4 : Discrimination Power Index of Test-Pack 2

LIST OF APPENDICES

- Appendix 1 : English Test-Pack Designed by English KKG of Ministry of Education and Culture Semarang (Test-Pack 1)
- Appendix 2 : English Test-Pack Designed by English KKG of Ministry of Religion Semarang (Test-Pack 2)
- Appendix 3 : Statistical Analysis of Multiple Choice Question of Test-Pack 1 using ITEMAN
- Appendix 4 : Statistical Analysis of Multiple Choice Question of Test-Pack 2 using ITEMAN
- Appendix 5 : Analysis Result of Short Answer Items test of Test Pack 1
- Appendix 6 : Analysis Result of Short Answer Items test of Test Pack 2
- Appendix 7 : Analysis Result of Essay Items test of Test Pack 1
- Appendix 8 : Analysis Result of Essay Items test of Test Pack 2
- Appendix 9 : Analysis Result of Validity, Level of Difficulty, Discrimination Power and Reliability of Short Item Test of Test Pack 1
- Appendix 10 : Analysis Result of Validity, Level of Difficulty, Discrimination Power and Reliability of Short Item Test of Test Pack 2
- Appendix 11 : Analysis Result of Validity, Level of Difficulty, Discrimination Power and Reliability of Essay Test Item Test of Test-Pack 1
- Appendix 12 : Analysis Result of Validity, Level of Difficulty, Discrimination Power and Reliability of Essay Test Item Test of Test-Pack 2
- Appendix 13 : Appropriateness to Curriculum of Test-Pack 1 and Test-Pack 2

Appendix 14 : The Characteristics of Good Test of Test Pack- 1

Appendix 15 : The Characteristics of Good Test of Test Pack- 2

THE VALIDITY, RELIABILITY, LEVEL OF DIFFICULTY AND APPROPRIATENESS OF CURRICULUM OF THE ENGLISH TEST

(A Comparative Study of The Quality of English Final Test of The First Semester Students Grade V Made By English KKG of Ministry of Education and Culture and Ministry of Religion Semarang)

Athiyah Salwa
13020210400002

Abstract

The main objective of this research is to present and compare the quality of two test-packs involving validity, reliability, level of difficulty, discrimination power, distractors' distribution and the appropriateness of curriculum and the characteristics of a good test. By conducting this research, the writer hopes the quality of test-packs that are used in the end of semester of elementary schools can be improved.

It studies the quality of the English test, especially English final test for the first semester students' grade V. This test was analyzed by descriptive comparative method with quantitative approach. Not only using quantitative approach, qualitative approach was also used to synchronize the tests with Standard and Basic Competence, and the characteristics of a good test (content validity). The test items used as the sample were English test-packs of the first semester students for Grade V of elementary schools designed by English KKG of Ministry Education and Culture and Ministry of Religion Semarang. The study only analyzed the Grade V of Elementary School just because of the limitation of the time of research.

In analyzing the data, the writer used several formulas to measure the tests' validity, reliability, level of difficulty, and discrimination power. She also used the ITEMAN program to measure distractors' distribution. The instruments used to analyze the data were curriculum checklist, observation checklist, test paper, and students' answer sheet.

The findings were in the form of index number of validity, reliability, level of difficulty, and discrimination power in the case of quantitative analysis. In qualitative analysis, the findings were in the form of percentage of test-items that fulfill the appropriateness of curriculum and some errors that exist in both test-packs. From the findings, the discussion came to the conclusion that the qualities of both test-packs are good in their quantitative aspects. The number of validity, reliability, difficulty index, and discrimination power of both test-packs are balances. However, in their qualitative aspects, test-pack 1 has better quality than test-pack 2. It is because the findings that there are some errors exist in test-pack 2. Thus, the writer suggests that test-makers of test-pack 2 have to be careful and notice the requirement of designing a good test in the next arrangement.

Keywords: Validity, Reliability, Level of Difficulty, and Discrimination Power, Appropriateness of Curriculum

VALIDITAS, RELIBILITAS, TINGKAT KESUKARAN DAN KECOCOKAN PADA KURIKULUM SOAL BAHASA INGGRIS

**(Studi Perbandingan Kualitas Soal Ulangan Akhir Semester I Kelas V yang
Disusun oleh KKG Bahasa Inggris Kementerian Pendidikan Dan Kebudayaan
Dan Kementerian Agama Kota Semarang)**

Athiyah Salwa
13020210400002

Intisari

Penelitian ini bertujuan untuk memaparkan dan membandingkan kualitas dua soal tes yang meliputi validitas, reliabilitas, tingkat kesukaran, daya pembeda, sebaran jawaban, dan kecocokan terhadap kurikulum dan kriteria soal yang baik. Melalui penelitian ini penulis berharap kualitas kedua soal yang digunakan di sekolah dasar di akhir semester pertama dapat ditingkatkan.

Penelitian ini menyelidiki tentang kualitas soal Bahasa Inggris khususnya yang digunakan pada semester pertama sekolah dasar kelas V. Soal Bahasa Inggris ini dianalisis menggunakan metode deskriptif komparatif dengan ancangan kuantitatif. Selain itu, penulis juga menggunakan ancangan kualitatif untuk memeriksa apakah soal tersebut sesuai dengan Standar Kompetensi dan Kompetensi Dasar pada kurikulum dan kriteria tes yang baik. Soal Bahasa Inggris yang digunakan sebagai sampel adalah Soal Bahasa Inggris Semester Pertama Kelas V Sekolah Dasar yang dibuat oleh KKG Bahasa Inggris Kementerian Pendidikan dan Kebudayaan dan Kementerian Agama Semarang. Penulis hanya menganalisa pada soal Bahasa Inggris Kelas V karena keterbatasan waktu.

Dalam menganalisis data, penulis menggunakan beberapa rumus untuk mengukur validitas, reliabilitas, tingkat kesukaran dan daya pembeda tes. Selain itu, untuk mengukur sebaran jawaban, penulis menggunakan aplikasi ITEMAN. Instrumen yang digunakan untuk menganalisis data berupa ceklist kurikulum, ceklist observasi, lembar soal, dan lembar jawaban siswa.

Hasil temuan berupa nilai indeks validitas, reliabilitas, tingkat kesukaran, dan daya pembeda dalam hal kuantitatif analisis. Sedangkan pada analisis kualitatif, hasil temuan berupa prosentase kecocokan soal pada kurikulum, dan beberapa kesalahan yang ada pada kedua tes.

Dari hasil temuan, dapat disimpulkan bahwa kualitas kedua soal baik dari segi kuantitatifnya. Nilai validitas, reliabilitas, tingkat kesukaran, dan daya pembeda keduanya seimbang. Namun, dari segi kualitatifnya, soal 1 lebih baik dari soal 2. Hal ini dikarenakan beberapa kesalahan yang ditemukan dalam soal tes 2 lebih banyak dari pada soal tes 1. Penulis menyarankan pembuat soal tes 2 harus berhati-hati dan memperhatikan ketentuan pembuatan soal yang baik pada penyusunan tes selanjutnya.

Kata Kunci: Validitas, Reliabilitas, Tingkat Kesukaran Soal, dan Daya Pembeda, Kecocokan pada Kurikulum

CHAPTER I

INTRODUCTION

1.1 Background of the Study

Evaluation in education grows more important nowadays. The aim of evaluation is to evaluate students' achievement and teachers' progress in teaching and learning process. Evaluation in education can be assumed as a formal and informal of examining students' achievement. Informal evaluation usually occurs by the time of teaching and learning process taking place. Teachers can evaluate the students' achievement by observing and making judgment based on students' performance during the process of teaching and learning. Yet, teachers cannot assume that students who never perform actively during the teaching and learning process do not understand the materials at all. It is because somehow students do not feel free to express their ideas. Thus, it needs a formal assessment to examine the students' understanding.

Teachers can do an evaluation by making an assessment. Evaluation can be done by making an assessment, but evaluation occurs in some ways by an observation or performance judgment during the process. Teachers, trainers, or education practitioners usually use the assessment to measure and analyze students' achievement.

To assess students' achievement of the material which has been taught to them, usually the teachers give their students some questions in the form

of a test. Teachers can conduct it after each chapter of the material is finished or in the end of semester. The test can be in the form of essay test in which students have to write the answer on some sentences. Besides, teachers can give the test in the form of multiple-choices to simply check students' achievement.

Testing language subject, in this case English, does not only examine the science and knowledge of the subject but also the skills of it. It is supported by Hughes (2005:2) who stated that "language ability is not easy to measure; we cannot expect a level of accuracy comparable to those measurements in the physical science". Thus, the language testing questions have to measure the learners' mastery of listening, speaking, reading and writing. Of course, the skills they have to master are in line with the students' level of education. It is for example in the level of senior high schools; the students should master at least two or three skills as a minimum requirement. It means that even though they are not able to speak or write English well at least they have to understand what they listen to or read.

In the level of elementary school, the students can be considered as mastering English reading skill when they can understand simple English sentence or text. The students of elementary school are said to be mastering English lessons when they are able to understand and make simple sentences in a school or class context either orally or in written. In other words, in elementary school level, the formal tests are usually only measuring students' achievement on reading and writing skills. The

achievements of listening and speaking skills are measured by the teachers during the process of teaching and learning process.

The formal tests used in Indonesia are usually the combination of multiple-choice and essay questions. Commonly, test-makers prefer to use multiple-choice question than essay test because it is effective, simple, and easy to score. Some formal tests like UAN or SNMPTN are in multiple-choice questions since they are given to a big number of test-takers. Yet, if a test is used in a certain condition like school or class context, teachers can combine different kinds of testing techniques such as combining multiple-choice question and essay test type. The combined test is usually used in summative test or final test. Unlike multiple-choice question, this test can measure students' ability in some skills of language. Teachers can evaluate students in several aspects but then again, it needs more time to score and analyze it than in multiple-choice question.

The combined test means a test that consists of multiple-choice question, essay test items, and sometimes short-answer items. The use of this test is to evaluate students' achievement and their ability to elaborate an idea. It is very good to be used in assessment process regarding it can measure students' whole knowledge about materials. That is why teachers in many formal Indonesian schools use this test as summative test. This kind of test is used in all levels of education, from Elementary School, Junior High School, until Senior High School. Unlike formative test which is designed and constructed by the teachers after each chapter of the material is finished,

the summative test is given to students in the end of semester. Thus, there are some themes of materials constructed in the test. Usually, it is designed by a group of teachers in one area or domain that is called KKG (Teachers Work Group) in the level of Elementary Schools and MGMP (Conference of Subject's Teachers) in the level of Junior and Senior High Schools.

In Indonesia, there are two ministries that have an authority to publish summative test used in schools. They are Ministry of Education and Culture and Ministry of Religion which manages Islamic based schools. This test is constructed by KKG of both ministries on each subject including English.. Since there are two institutions publishing the test, there are two versions of test-packs given to the students as final semester test. Considering that both test-packs are made by different institution; they may have different characteristics and qualities even though they are used in the same grade and level of education.

Considering the importance of measuring and examining students' achievement, it is important to the teachers to design a good test. A good test can present students' achievement well. A test can be said as a good test if it fulfills several requirements of a good test, both statistically and non-statistically. By presenting both aspects, we can see then the quality of the test in order to decide whether the test is good enough to be used or not. If it does not fulfill the requirements of a good test, test-makers should redesign and rearrange it. A problem arises when there are two different test packs of the same grade of each education level. A question comes up whether or not

the test-packs organized by Ministry of Education and Culture has the same quality and characteristics with the one arranged by Ministry of Religion. If there are some differences, it is unfair to use those tests. Another case is that one of the test packs may not be appropriate with instructional material, in this case Standard and Basic Competence.

Based on the explanation above, the writer was interested in conducting a research that studies the comparison of the qualities of both test-packs. The writer then formulated the title of this study as “The Validity, Reliability, Level of Difficulty and Appropriateness to Curriculum of English Final Test”. This study uses the sample of English test of the first semester students’ grade V academic year of 2011/2012 made by English KKG of Ministry of Education and Culture and Ministry of Religion of Semarang”. This title is made by the reason that quality of a test can be gained by analyzing its statistical quality, such as its validity, reliability, level of difficulty and discrimination power, and non-statistical quality, such as its appropriateness to curriculum.

1.2 Identification of the Problems

English Final Test of the first semester used in Elementary Schools of Semarang is made by two different institutions. Those two institutions are English KKG of Ministry of Education and Culture and Ministry of Religion.

The writer found that English test-pack designed by Ministry of Religion has some errors and may be lack of correction. There are some

mistype, misspelling, and grammar errors on its test-items. She assumed that English test-pack designed by Ministry of Education and Culture is better than test-pack designed by Ministry of Religion. A problem arises if the test-pack designed by Ministry of Religion is used again in the next semester before it is revised and analyzed. The errors would exist again in the same test-pack.

From this problem, the writer wanted to compare and describe the quality of both tests in case of their quantitative and qualitative aspects. In quantitative aspect, the validity, reliability, level of difficulty, and discrimination power of a test-pack are measured and analyzed. Meanwhile, their qualitative aspect is measured by looking for their appropriateness to curriculum. Therefore, the quality of both test-packs can be known and fixed accurately based on the analysis.

1.3 Statements of the Problems

In order not to discuss something irrelevant the writer has limited the discussion by presenting and focusing her attention to the following problems:

- 1.3.1 To what extent the quality of the English Final test of the first semester students made by English KKG of the Ministry of Education and Culture and Ministry of Religion of Semarang in terms of their validity, reliability, difficulty level, discrimination power, and item distractors?

- 1.3.2 To what extent the appropriateness of those test items to the curriculum (Standard Competence and Basic Competence), and characteristics of a good test?
- 1.3.3 Is there any significant difference between tests items made by English KKG of Ministry of Education and Culture and Ministry of Religion of Semarang?

1.4 Objectives of the Study

Based on the formulated problems above, this study has several objectives elaborated as follows:

- 1.4.1 To present the quality of the English Final test of the first semester students made by English KKG of the Ministry of Education and Culture and Ministry of Religion of Semarang in terms of their validity, reliability, difficulty level, discrimination power, and item distractors?
- 1.4.2 To present the appropriateness of those test items of the curriculum (Standard Competence and Basic Competence), and the characteristics of a good test?
- 1.4.3 To find out the significant difference between the tests items made by English KKG of Ministry of Education and Culture and Ministry of Religion of Semarang?

1.5 Significances of the Study

Related to the objectives of the study, this analysis is intended to see some advantages as elaborated in some paragraphs below. There are three major significances that this study wants to contribute.

The first one is theoretical significance. This study may give basic understanding to the teachers, educators, trainers, and others that assessment and evaluation cannot be made and assumed only by basing on students or one's outer performance or guessing in some cases. They should know that the test items should be made to evaluate students' understanding and ability. The tests are also useful to develop their professionalism as being an educator.

The second one is practical significances. This study is beneficial for the test makers as additional reference in constructing and analyzing test items and their procedures.

The last one is pedagogical significance. This study provides English teachers especially elementary schools' teachers with some meaningful and useful information of efficient class discussion of the test result, the general improvement of classroom instruction, evaluation in teaching learning process, and improvement in test making.

1.6 Scope of the Study

This study is quantitative and qualitative research. It studies the quality of the English test, especially English final test for the first semester

students' grade V. This test was analyzed by using descriptive comparative method with quantitative and qualitative approach.

The test items used here are English test-packs in final test of first semester students for Grade V of Elementary School. The study only analyzed the Grade V of Elementary School just because of the limitation of the time of research.

1.7 Definition of the Key Terms

There are several key terms that are used in this study. They are Validity, Reliability, Level of Difficulty, and Discriminating Power. They are defined in some paragraphs below:

- 1) The Validity of a test represents the extent to which a test measures what it purports to measures (Tuckman, 1978:163).
- 2) Reliability is consistency of measures across different conditions in the measurement procedures (Bachman, 2004: 153).
- 3) Level of difficulty (Item Facility) is the extent to which an item is easy or difficult for the proposed group of test-takers (Brown, 2004:58). Gronlund (1993: 103) states that difficulty level of an item in a test is the percentage of students who answer test items correctly.
- 4) Discriminating Power (Item Discrimination) is the ability of the test items measures the better and poorer examinees of items (Remmers, Gage and Rummel, 1967: 268). In the same context, Blood and Budd (1972) defined the index of discrimination as the ability of an item on

the basis of which the discrimination is made between superiors and inferiors.

1.8 Research Hypothesis

Hatch (1982: 3) states that hypothesis is a tentative statement about the outcome of the research. The general definition of it can be said as pre-assumption of the researcher about the product of the study. In this research, the hypothesis (Ha) is that the quality of English Final test of first semester of Fifth Grade of Elementary School constructed by KKG of English of Ministry of Education and Culture and Ministry of Religion has the same quality in case their quantitative and qualitative aspects. Null hypothesis (Ho) of this study is that qualities of both test-packs are different.

1.9 Outlines of the Study Report

In order to make the readers become easier in understanding this study report, the writer is going to organize this research paper as follows:

Chapter I is Introduction. It includes the explanation about the background of the study, identifications of the problem, statements of the problem, objectives of the study, significances of the study, underlying theories, scope of the research, research method, definition of key terms, and the outline of the study report.

Chapter II presents review of related literature that consists of the definition and Previous Studies, School-Based Curriculum (KTSP),

Teachers Work Group (KKG), Language Testing and Assessment, Types of Assessment and Testing, Characteristics of a Good Test, and Item Analysis.

Chapter III deals with research method. It presents research design, population and sample, research instrument, method of collecting the data, instruments, and method of analyzing the data.

Chapter IV presents research findings and discussion. It consists of description of the findings and discussion of it.

Chapter V as the end of the discussion includes the conclusions and suggestions.

CHAPTER II

REVIEW OF THE RELATED LITERATURE

This chapter presents some references related to this study. They are the explanation of the Previous Studies, School-Based Curriculum (KTSP), Teachers Work Group (KKG), Language Testing and Assessment, Types of Assessment and Testing, The Characteristics of a Good Test, and Item Analysis.

2.1 Previous Studies

This research refers to the previous study by Ema Rahmatun Nafsah (2011) entitled “An Analysis of English Multiple Choice Question (MCQ) Test of 7th grade at SMP BUANA Waru Sidoarjo” and Hastuti Handayani (2009) entitled “An Analysis of English National Final Exam (UAN) For Junior High School viewed from School-Based Curriculum (KTSP)”.

Nafsah examined English Multiple Choice Question that was constructed by English teacher in a school. Her research is descriptive qualitative research. She tried to know the quality of the test that was independently designed by the English teacher. The source of the data in her study is English final test items designed by the teachers, the students’ answer sheet, and the students’ scores of 7th grade students in SMP BUANA especially for 7B, 7D, and 7E. Those three classes are the sample of her study because she took the data randomly. The result of her study leads to the

conclusion that English Multiple Choice Questions (MCQ) Test constructed by an English teacher of 7th grade in SMP BUANA Waru Sidoarjo has good test based on the characteristics of a good test, good face validity and high content validity, high reliability, good index of difficulty but poor index of discrimination.

In line with the analysis of English test-pack, Handayani (2009) conducted an analysis about English formal test entitled “An Analysis of English National Final Exam (UAN) For Junior High School viewed from School-Based Curriculum (KTSP)”. Her research is descriptive and content analysis. She investigated the appropriateness of English test-packs used in National Final Exam (UAN) to the School-Based Curriculum (KTSP). The main data of this research are material of English UAN for SMP/MTs academic year of 2006/2007 and 2007/2008. The units of analysis are sentences and texts. In analyzing the data, she used some instruments. They are matrix of competence standard and basic competence (curriculum) which covers discourse competence in reading, writing, speaking, and listening skill. The result of this study came to an end by the conclusion that most of materials (test-items) of the English National Final Examination academic year of 2006/2007 and 2007/2008 match with Content Standard and Competencies of English syllabus for SMP in Semarang. Even though there are five items of the English UAN academic year of 2006/2007, all in all the materials contain competencies for all skills, whereas, English UAN academic year of 2007/2008 only contains reading and writing skill only. As

the previous test-packs, it matches to the syllabus and the content standard. The mistake of English UAN academic year of 2006/2007 did not happen again in this test-pack.

Related to the previous studies, this research was conducted to complete and improve those two previous researches. The writer combined two methods and analysis instrument of both previous studies. It used English first semester test, as a formal test like English UAN. The analysis of it involves item analysis such as validity, reliability, level of difficulty, discrimination power, and appropriateness to curriculum as Nafsah's thesis.

2.2 School-Based Curriculum (KTSP)

Curriculum is a document of an official nature, published by a leading or central education authority in order to serve as a framework or a set of guidelines for the teaching of a subject area in a broad varied context (Celce-Murcia, 2000). A curriculum in a school refers to the whole body of knowledge that children acquire in school (Richards, 2001:39). More specific, BSNP (*Badan Standar Nasional Pengembangan*) (2006:1751) defines it as a set of plan and arrangement of objective, content, and lesson material, and also manner that is used as the guidance of learning activities to achieve the aim of education. In short, we can say that curriculum is the fundamental guidelines for teachers to reach the aims of education in school. It is a ground-base that teachers should know in conducting teaching learning process.

School-Based curriculum is a revised-edition of curriculum of 2004 which is in Bahasa Indonesia stated as *Kurikulum Berbasis Kompetensi* (Competence-Based Curriculum). This curriculum is firstly established in 2006. It is the way in which a school can create and make policy and rule of its educational programs. Teachers can create their own syllabus, teaching-learning processes, and learning goals that are appropriate for students in their school.

The content of both KBK and KTSP are not different. KBK is designed and established by official institution in this case Education and Culture Ministry, while KTSP is created by the school itself based on KTSP Arrangement Guide established by BSNP (Muslich, 2009:17)

KTSP is an operational curriculum that is arranged and applied in every educational unit (Muslich, 2008:10). It is created based on the school's need and condition. In this case, schools in big city may have different curriculum from the schools in a small city. The arrangement of the content itself is regarding to the cultural and social condition of the students. Thus, the students in different places and areas have their own learning achievement that appropriate to their natural life. Even though based on Government Rule 19, 2005 about Education National Standard, every school is mandated to develop KTSP based in Passing Competence Standard (SKL), and Content Standard (SI) and based on the guidance arranged by Education National Standard Board (BSNP). A school is called having ability to arrange and develop KTSP if it tries to apply Curriculum of 2004 on its institutions.

Based on the Rule of Minister of National Education number 24, 2006, the arrangements of KTSP involves teachers, employees, and also School Committee with the hope that KTSP will reflect the aspiration of people, environment situation and condition, and the people's need. Because of that this curriculum is more democratic than the previous curriculum. The writer presented competence standard and basic competence of English Lesson grade V semester I that related to this study. It can be seen in the table below:

Table 2.1: Competence Standard and Basic Competence of Fifth Grade of Elementary School

Competence Standard	Basic Competence	Indicator
1. Listening Students are able to understand very simple instruction with an action in school context	1.1 Students are able to respond very simple instruction with logical action in class and school context	<ul style="list-style-type: none"> • Students are able to complete a sentence in form of Present Continuous Tense. • Students are able to mention imperative sentence. • Students are able to answer a question using a sentence in form of Simple Present Tense.
	1.2 Students are able to respond very simple instruction verbally	<ul style="list-style-type: none"> • Students are able to make conversation dialogue. • Students are able to decide correct or incorrect statement based on a text. • Students are able to make a

		<p>conversation about ordering a menu in a restaurant.</p> <ul style="list-style-type: none"> • Students are able to classify information on a text into a table. • Students are able to answer a question about dancing from several countries.
<p>2. Speaking Students are able to express very simple instruction and information in school context</p>	<p>2.1 Students are able to make a very simple conversation that follow logical action with speech act ; give an example to do an action, give a command, and give an instruction</p>	<ul style="list-style-type: none"> • Students are able to read a story in form of Simple Present Tense. • Students are able to answer a question using a sentence in form of Simple Present Tense.
	<p>2.2 Students are able to make a very simple conversation to ask and or give something logically involve speech act , asking and give a help, asking and giving something</p>	<ul style="list-style-type: none"> • Students are able to pronounce a word can correctly in a simple sentence. • Students are able to mention things in a medicine box. • Students are able to identify a picture related to certain sentence. • Students are able to differentiate the use of how many and how much in a question.
	<p>2.3 Students are able to ask and give information</p>	<ul style="list-style-type: none"> • Students are able to mention the name of several shapes.

	involve speech act; introducing, inviting, asking and giving permission, agreeing and disagreeing, and prohibiting	<ul style="list-style-type: none"> Students are able to differentiate <i>simple present verb and simple past verb.</i> Students are able to tell their past activities.
	2.4 Students are able to express politeness using expression: <i>Do you mind</i> and <i>Shall we...</i>	<ul style="list-style-type: none"> Students are able to mention the name of <i>musical instruments.</i> Students are able to clarify information from a text into a table. Students are able to answer a question about <i>dancing from several countries.</i>
3. Reading Students are able to understand English written texts and descriptive text using picture in school context	3.1 Students are able to read aloud with stress and intonation correctly involve words, phrases, and simple sentence.	<ul style="list-style-type: none"> Students are able to read a story in form of Simple present tense.
	3.2 Students are able to understand simple sentence, written messages, and descriptive txt using picture accurately	<ul style="list-style-type: none"> Students are able to read a story based in a text. Students are able to read a story in form of <i>Simple past tense.</i> Students are able to <i>mention time.</i> Students are able to read a story in form of comic. Students are able to read a poem.

		<ul style="list-style-type: none"> Students are able to read a short story.
4. Writing Students are able to spell and rewrite simple sentence in school context	4.1 Students are able to spell simple sentence accurately and correctly	<ul style="list-style-type: none"> Students are able to complete a sentence in form of Present Continuous tense. Students are able to complete a sentence with simple past verbs.
	4.2 Students are able to rewrite and write simple sentence accurately and correctly; such as compliment, felicitation, invitation, and gratitutation	<ul style="list-style-type: none"> Students are able to identify a picture related to certain sentence Students are able to decide correct or incorrect statement based on a text Students are able to make a conversation about ordering a menu in a restaurant. Students are able to identify some types of food. Students are able to use <i>adverb of manner</i> in a sentence. Students are able to answer <i>mathematical questions</i>.

2.3 Teachers Work Group (KKG)

Teachers work group (KKG) is a group of teachers that is organized to improve and develop teachers' professionalism especially the teachers of

Elementary Schools. In the level of Elementary School it is called as KKG and MGMP for the level of Junior and Senior high school. Regarding that the goal of this association is to maintain and develop teachers' professionalism, it has some goals and programs.

Based on *Standar Pengembangan KKG dan MGMP* (2008:4), there are some goals of KKG/ MGMP. They are:

- 1) To improve teachers' knowledge and concept of teaching and learning material substances, learning methods, and to maximize learning media.
- 2) To give a place and media for teachers as members of KKG/ MGMP to share their experience, ask and give for solution.
- 3) To improve teachers' skill, ability, and competencies in teaching and learning process through the activities of KKG/MGMP.
- 4) To improve education and learning process quality as reflected in students' achievement improvement.

The programs of KKG/MGMP are arranged by its members and acknowledged by Principal Work Group (KKKS)/ Principal Work Conference (MKKS). It is legalized by Chairman of Education Official.

There are two main programs of KKG/MGMP. They are routine programs and development programs. The routine programs consist of learning problem discussion, arrangement of lesson plans, syllabus, and semester programs, curriculum analysis, arrangement of learning evaluation instrument, and Final Examination preparation. The development programs are optional to be done. It can be in the form of research, seminar and

training, journal arrangement, Peer Coaching, Lesson Study, and Professional Learning Community.

The members of KKG are classroom teachers, and subject teachers from 8-10 schools. They consist of chairman, secretary, treasurer, and the members. There is no difference between KKG of SD and MI. It means that KKG of SD has the same programs, goals, and management as KKG of MI does. Thus, it can be assumed that there are no differences between test-packs designed by both KKG of different ministries.

2.4 Language Testing and Assessment

A test is a method of measuring a person's ability, knowledge or performance in a given domain (Brown, 2004:3). By this definition, Brown wants to highlight on the term testing as a way or method in which people's intelligence and achievement are being explored. Testing becomes the important method to check many requirements or competency in some fields like medicine, law, sport, and government. Yet, in teaching and learning process, the term testing is little bit different from those kinds of test. Related to the term of testing, people commonly think that assessment is the same method as testing. They are still confused and consider that testing and assessment are synonymous.

Alderson and others have argued that "testers have long been concerned with matters of fairness and that striving for fairness is an aspect of ethical behavior, others have separated the issue of ethics from validity, as an essential part of the professionalizing of language testing as a discipline" (in

Davies, 1997). In short, it can be said that test is a part of assessment so that assessment is wider than test itself. Assessment can be understood as a part of teaching and learning process. Testing and assessment are two methods that must be used and implied in teaching.

There are several principles of language assessment as Brown (2004: 19-28) stated that are practicality, reliability, validity, authenticity and washback. Yet, in this study only some principles that are examined more detail. They are items of analysis consisting of the validity, reliability, level of difficulty and item discrimination. They are explained more in paragraphs below.

2.5 Types of Assessment and Testing

In order to know more about assessment, in this sub chapter the writer wanted to explain about type and form of assessment. There are two types of assessment, informal and formal assessment (Brown, 2004:5). Informal assessment can take a number of forms starting from incidental, unplanned comments and responses, along with coaching and other impromptu feedback to the student (Brown, 2004:5). In this type of assessment, teachers record students' achievement by some techniques that are not systematically made. Teachers can memorize what students do in the classroom based on their learning activity. Whereas, formal assessment are exercises or procedures specifically designed to tap into a storehouse of skills and knowledge (Brown, 2004:5). Different from informal assessment, this type of assessment is intentionally made by teacher to get students' score to know their

achievement. This assessment is done by teachers by making standard and official based on the rule.

Two functions of assessment that usually occur in the classroom based are formative and summative assessment (Brown, 2004:6). Formative assessment intends to evaluate students in the process of forming their competencies and skills with the goal of helping them to continue that growth process (Brown, 2004:6). This formative assessment usually occurs during teaching and learning process in the classroom. It is done by the teachers to know directly students' achievement. This assessment is conducted to build and grow up students understanding and skills during the process. Assessment is formative when teachers use it to check on the progress of their students, to see how they have mastered what they should have learned, and then use this information to modify their future teaching plans (Hughes, 2005:5). Summative assessment, then, aims to measure, or summarize, what students have grasped, and typically occurs at the end of a course or unit of instruction (Brown, 2004:6). It is used in the end of the term, semester, or year in order to measure what have been achieved by pupils. This type of assessment is used by the teachers to measure and evaluate what students achieved during the process of teaching and learning in classroom. Final exams are the example of this test. In short, formative assessment is done in the middle of the semester in the process of teaching and learning, but summative is done in the end of the semester. The object of this study is final

test of first semester, so this kind of test is formal assessment with the function of summative assessment.

In Indonesia, usually a final semester test-packs consist of three parts of items. They are, first, multiple choice items, the next is short-answer question, and the last is essay items. Every item has different definitions and characteristics. There are some different formulas and measurement that can be used. To know more about the characteristics of each item, next sub-chapter below explains more about them.

2.5.1 Multiple-Choice Test

Multiple-choice Question test is the simplest test technique commonly used by test-makers. It can be used any condition and situation, in any level or degree of education. Actually, its simplicity relies on its scoring and answering. It is supported by Hughes (2005:75) who states the most obvious advantage of multiple-choice is that scoring can be perfectly reliable. In line with Hughes, Valette (1967:6) states that scoring in multiple choice techniques is rapid and economical. And it is designed to elicit specific responses from the student.

Yet, designing multiple-choice question is more complicated than essay items. According to Brown (2004:55) multiple-choice items which may appear to be the simplest kind of item to construct are extremely difficult to design correctly. Multiple-choice items take many forms, but their basic structure is that it has stems or the question itself, and a number of options-one which is correct, the others being distractors (Hughes, 2005:75).

In another case, Hughes states number of weaknesses of multiple-choice items (Hughes, 2005:76-78). Multiple-choice questions only recognition of knowledge. They make test takers can only guess to come with correct answer, and cheat easily. The technique severely restricts what can be tested. It is very difficult to write successful items and the answer is restricted by the optional answer. In this case, test-takers can not elaborate their answer and understanding of the material because the answer is limited only by an optional answer.

Multiple-choice comes to be the first part of test packs faced by test-takers. When we want to analyze this item we can use statistical analysis as stated in the next chapter. Since there is only one right answer, the score can very rapidly mark an item as correct and incorrect (Valette, 1967:6). Thus, we can use simple codes to present the answer of test-takers. Score 1 presents correct answer chosen by students, and 0 presents wrong answer. If students choose a correct answer, we can note it by 1. And vice versa, if test-takers answer with wrong answer we note it with number 0.

2.5.2 Short-Answer Items

After test-takers have already answered the multiple choice items in first chapter of test-packs, in the next chapter they have to answer on short-answer items. The question is just the same, but in these items students are not given an optional answer. The answers are usually only one or two words. Those answers should be exactly correct, but the exactly correct answer usually occurs in only listening and reading tests (Hughes, 1989:79).

Regarding that English first semester test contains reading and writing skills, student's answer of this items especially on reading skill should exactly correct.

Short-answer items deal with measurement of students' knowledge acquisition and comprehension. It has two choices or formats, free and fixed. Basically, there are two basic free formats. They are unstructured format and fill-in or completion format. Fixed choice format, then, consists of true-false, other two-choice, multiple-choice and matching (Tuckman, 1975:77). Short-answer items in English final semester test-packs used in this study here are the items in which students should answer by writing down the answer in a short and brief sentence. They are different from essay-test items. In essay-test items, students should explore and elaborate their answer. For example, if the question is about structure and grammar, usually students should fill in the blank with a complete sentence. Yet, in short-answer items what students should answer are usually not more than two or three words. As Valette (1967:8) states that this item may require one-word answer, such as brief responses to questions, or the filling in of missing elements.

In the short-answer items, the true answer has been determined by teachers so that students can not elaborate their answer. Both free choice and fixed choice items have previously determined correct response. In this formats, basically, measurement involves asking students a question that requires that they state or name the specific information or knowledge (Tuckman, 1975:77).

Sometimes, in short-answer items are in form of unstructured and completion/ fill-in format. In unstructured format, students can answer by a word, phrase or number. While in completion or fill-in format, students must construct their own response rather than choose an optional answer.

In order to assure to the objective nature of short-answer items, teacher must prepare a scoring system in advance (Valette, 1967:8). Teacher should give credit score to students' answer for misspelling of the word given. But since in short answer usually the answer is only one word, we can use the credit point the same as multiple choice. We can use the score 1 to presents students chosen correct answer and number 0 that presents incorrect answer. We only have to mark as 1 and 0 because the answer has been determined by test-makers and there is no optional answer for test-takers.

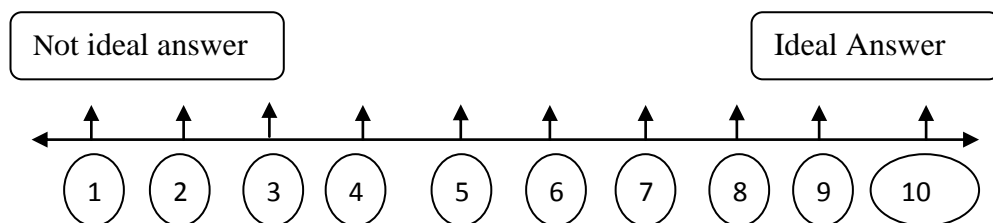
2.5.3 Essay Test Items

In English final test of elementary school, beside multiple choice and short-answer items, there is one more test technique that is served to the test-takers in final semester test-packs. It is essay test. Different from short-answer items, essay test needs longer sentence to answer it. While short answer is the continuity of multiple choice items, essay-test items involve deep thinking about test-takers knowledge and understanding on material. In language testing, it may include in students understanding on language structure and culture. It is supported by what Tuckman (1975:111) stated that "Essay items provide test-takers with the opportunity to structure and compose their own responses within relatively broad limits enable them to

demonstrate their ability to apply knowledge and to analyze, to synthesize, and to evaluate new information in the light of their knowledge.”

The scoring system of this item will be very different from scoring objectives items or multiple-choice. In objective items, the score of each number is exact and all the same from number to number. Whereas, in essay items, what we should do, first, is determining the ideal answer even though no correct and wrong answer at all. The ideal answer then should be scored as the highest score. The far answers of students go beyond it will be the lowest score it is. Teachers then should create interval scale to score the highest and the lowest one on each item. Interval scale will be going like picture below:

Diagram 2.1: Interval Scale for Essay-test items Scoring



The interval scale then can be used to measure how far students understand the material. If the students get higher score, it means that they understand more on the material. Teachers have an authority to determine interval scale number between ideal and not-ideal answer. It can be a scale from 0 until 10 like the scale above, or 0 until 3 or 5 based on their preferences. It may be decided by calculating every score of every item, from the multiple-choice questions, short-answer items, and the last one is essay-test items.

2.6 Characteristics of a Good Test

Based on Zulaiha (2008:1) on her book *Manual Test Analysis*, to get to know about the characteristics of test items, we should do an analysis on their quantitative and qualitative aspects. Qualitative analysis is used to know whether test-items will function properly or not to test students, while, quantitative analysis is used to know the use of test-items after given to the students.

In quantitative analysis, we have certain formulas to measure the statistical quality of multiple-choice question, short answer and essay items. Yet, in qualitative analysis there are several characteristics that have to be fulfilled in order to be said as functional items. Based on Zulaiha (2008: 2), multiple-choice items are good if they fulfill the characteristics as follow:

- 1) There should be one correct answer on each question
- 2) Only one feature at the time should be tested
- 3) Each option should be grammatically correct when placed in stems
- 4) It should be efficient in using word, phrase, and sentences.
- 5) The optional answer should be chronologically stated.
- 6) It should not be dependent on other question
- 7) The stems should not give clues or question to other question
- 8) All multiple choice items should be at a level appropriate to the proficiency level of education
- 9) Pictures, Graphics, tables, and diagrams should be clear and in function

- 10) The questions, statements, and spelling should be grammatically correct and clear.

Almost the same as the characteristics of good multiple-choice items, short-answer items and essay items have their own characteristics. It is a little bit different from multiple-choice because short answer and essay test items do not have optional answer. Based on Zulaiha (2008: 25-26), they are:

- 11) The limitation of the question and answer should be clear
- 12) The questions should be at a level appropriate to the proficiency level of education
- 13) The questions, statements, and spelling should be grammatically correct and clear
- 14) It should be efficient in using word, phrase, and sentences.
- 15) It should not be dependent to other question
- 16) The instruction of answering the items should be clearly stated
- 17) Pictures, Graphics, tables, and diagrams should be clear and in function

2.7 Item Analysis

Item Analysis is related to the several items of statistical analysis in analyzing characteristics and features of a test. They consist of validity, reliability, level of difficulty, discriminating power, and distribution of distractors.

2.7.1 Validity

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and action based on test scores or other modes of assessment. (Bachman, 2004:259).

The expert should look into whether the test content is representative of the skills that are supposed to be measured. This involves looking into the consistency between the syllabus content, the test objective and the test contents. If the test contents cover the test objectives, which in turn are representatives of the syllabus, it could be said that the test possesses content validity (Brown, 2002: 23-24). Brown's idea is supported by Hughes (2005:26), who stated that a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc which it is meant to be concerned. It means that a test will have content validity if the test-items are appropriate to what teachers want to measure.

To measure the validity of the test-items, the writer used the formulas of *product moment* below:

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

(Bachman, 2004:86 and Tuckman, 1978: 163)

For the detail explanation about the formula, it can be seen in the chapter III.

2.7.2 Reliability

Reliability refers to the consistency of test result. Reliable here means that a test must rely and fit on several aspects in conducting the test itself. A test should be reliable toward students. Bachman (2004: 153) states that reliability is consistency of measures across different conditions in the measurement procedures. Test administration must be consistent by which a test can be said as well-organized test. In vice versa, bad administration and unplanned arrangements of a test can make it does not work in measuring students' accomplishment. The writer used Alpha formula below to measure reliability of short answer and essay-test items only. Multiple-choice question items were already measured automatically by using ITEMAN program. The formula is:

$$r_{11} = \left(\frac{n}{n-1} \right) \left(1 - \frac{\sum \sigma_1^2}{\sigma_1^2} \right)$$

(Tuckman, 1978:163)

2.7.3 Level of Difficulty

A good test is a test which is not too easy or vice versa too difficult to students. It should give optional answer that can be chosen by students and not too far by the key answer. Very easy items are to build in some affective feelings of “success” among lower ability students and to serve as warm up items, and very difficult items can provide a challenge to the highest-ability students (Brown, 2004:59). It makes students know and record the characteristics of teacher's test if the test given always comes to them too

easy and difficult. Thus, the test should be standard and fulfill the characteristics of a good test. The number that shows the level difficulty of a test can be said as difficulty index (Arikunto, 2006:207). In this index there are minimum and maximum scores. The lower index of a test, the more difficult the test is. And vice versa, the higher the test, the easier it is.

There are some factors that every test constructors must consider in constructing difficulty level of test items. Mehren and Lehmen (1984) point out that the concept of difficulty or the decision of how difficult the test should depends on variety factors, notably 1) the purpose of the test, 2) ability level of the students, and 3) the age of grade.

The formula that can be used to measure it is:

$$IF = \frac{B}{JS} \quad (\text{Brown, 2004:59})$$

Another formula for measuring item difficulty (P-value) given by Gronlund, (1993: 103) and Garrett (1981:363) is below:

$$P = \frac{R}{N} \times 100$$

Where

P = the percentage of examinees who answered items correctly.

R = the number of examinees who answered items correctly.

N = total number of examinees who tried the items.

In measuring level of difficulty of an essay tests or short answer items, the writer used the different formula test below:

$$P = \frac{Mean}{\max imumscore} \quad (Zulaiha, 2008: 34)$$

2.7.4 Discriminating Power (Item Discrimination)

It is the extent to which an item differentiates between high and low-ability test-takers. Discrimination is important because if the test-items can discriminate more, they will be more reliable (Hughes, 2005:226). It can be defined also as the ability of a test to separate master students and non-master students (Arikunto, 2006:211). A master student is a student with higher scores of test, and a non-master student is a student with lower scores on the test given. The same as the term of difficulty level, discrimination has discrimination index. It is an indicator of how well an item discriminates between weak candidates and strong candidates (Hughes, 2005:226). This index is used to measure to the ability of a test in discriminating the upper and lower group of students. Upper students are students who answer with true answer, and lower group are students with false answer. In this index, it has negative point. Different from difficulty index, the negative index of discrimination power shows that the questions identify high group students as poor students and low group students as smart students. A good question is a question that can be answered by upper group and cannot be answered with true answer by lower group.

The categorizing of index of difficulty is divided into five types. They are too difficult, difficult, sufficient, easy, and too easy test-items. Its categorizing is based on the standard stated by Brown (2004:59), Arikunto (2006:208) that a test items is called too difficult if the number of P (index of difficult) is 0.00. Next, a test item is called difficult if the index is between 0.00-0.300. A test item is in range of sufficient if the index of difficulty is between 0.30-0.70. Then, it is called easy test if the index is between 0.70-1.00. It is called too easy if the number of P is equivalent to 1.00. The appropriate test item will generally have P that range from 0.15 to 0.85. (Brown, 2004:59)

An item will have poor index difficulty if it cannot differentiate between smart students and poor students. It happens if smart students and poor students have the same score on the same item. Conversely, an item that garners correct responses from most the high-ability group and incorrect responses from most of the low ability group has good discrimination power (Brown, 2004:59).

The formula that can be used to measure the discrimination power of multiple-choice test items is:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} \quad (\text{Brown, 2004:59})$$

Another version stated by Gronlund, (1993: 103) and (Ebel and Frisbie, 1991:231) is:

$$D = \frac{R_v}{N_v} - \frac{R_L}{N_L}$$

The same as level of difficulty, discrimination power also has different formula for essay test. It is because in essay test, each item of tests has highest and lowest score. To measure this, we can use the formula below:

$$D = \frac{MeanA - MeanB}{\max imumscore} \quad (Zulaiha, 2008: 34)$$

2.7.5 Answer of Questions Form (Item Distractors)

In addition to calculating discrimination indices and facility values, it is necessary to analyze the performance of distractors (Hughes, 2005:228). It is defined as the distribution of testee in choosing the optional answer (distractors) in multiple choice questions (Arikunto, 2006:219). This item is as important as the other items considering that in view of nearly 50 years of research that shows that there is a relationship between the distractors students choose and total test score (Nurulia, 2010:57).

It can be obtained by calculating the number of testee in choosing the distractors. We can calculate this form by seeing the answer form done by students. The distractors are good if chosen by minimum 5% of the number of test takers. One way to study responses to distractors is with frequency table that tells us the proportion of students who selected a given distractor. Remove or replace distractors selected by few or no students because students find them to be implausible (Nurulia, 2010:57). Distractors that are not chosen by any examinees should be replaced or removed. Distractors that do not work for example are chosen by very few test-takers should be replacing

by better ones, or the item should be otherwise modified or dropped (Hughes, 2005:228). They are not contributing the test's ability to discriminate the good students from the poor students (Nurulia, 2010:57). They should be discarded because they are chosen by very few test-takers from both groups. It means that they cannot function properly.

CHAPTER III

RESEARCH METHOD

This chapter consists of five sub chapters. They are research design, population and sample, research instruments, method of collecting data, and method of analyzing data.

3.1 Research Design

The research design used in this study was descriptive comparative with quantitative and qualitative approach. This study was descriptive because its aim is to present and describe the quality of the English test-packs. It was comparative since it used two samples for its data analysis and the writer compared those test-packs one to another to see whether there was difference between them or not.

Quantitative approach was used to measure the tests' validity, reliability, difficulty level, and discrimination power. To measure those items, several formulas were used. They are explained more detail in the next sub chapter. In addition, the qualitative approach was used to check whether or not the test items were appropriate with Standard and Basic Competence and fulfillment of the characteristics of a good test.

3.2 Population and Sample

The population of this study was English final semester test used in Elementary Schools in Semarang. The samples of them were English Final Test of the First Semester Students Grade V. There were two test-packs used. The first one was a test-pack designed by English KKG of Ministry of Education and Culture (in this study it is called as Test-pack 1), and the other one was designed by Ministry of Religion Semarang (in this study it is called as Test pack 2).

Those two test-packs were given as experiment testing to the students of two schools. They were SDIT Al Kamilah under the regulation of Ministry of Education and Culture and MI “Darus Sa’adah” under the management of Ministry of Religion. The goal of trying out both test-packs into two different schools was to know students’ score and answer to be used as the data for quantitative analysis.

3.3 Research Instruments

To collect the data needed for this study, the writer used four forms of instruments. They were curriculum checklist, characteristics of a good test checklist, question test paper, and students’ answer sheet. The detail explanation of each instrument can be seen as follows:

3.3.1 Curriculum Checklist

This instrument was used to know the appropriateness of the test-items within the curriculum and to know how far the test-items have fulfilled the instructional materials.

3.3.2 Characteristics of a Good Test checklist

It was used to identify some characteristics of English Final test paper. The writer examined test-items and analyzed them whether they have fulfilled the characteristics of a good test or not. With this instrument, the researcher got the qualitative data to answer the statements of problem in number 2.

3.3.3 Paper Test Question

It consists of multiple choice, short-answer items, and essay items. The two test packs were taken from English Final Test used by SDIT Al Kamilah Semarang and MI Darus Sa'adah Semarang. Each of the test packs was given into one class of grade V of SDIT Al Kamila Semarang and MI Darus Sa'adah. These two test-packs were delivered from different institution. The one used in SDIT Al Kamila was made by English KKG of Ministry of Education and Culture Semarang. The other used in MI Darus Sa'adah was made by English KKG of Ministry of Religion Semarang.

Those two test-packs then were matched with curriculum or instructional material to see the quality of both test-packs in case of their qualitative aspect. After they had matched, the writer drew an explanation about their quality to answer the question problem no 2.

3.3.4 Students Answer Sheet Paper

Besides using the main instruments, the two test-packs, the writer also used students' answer sheet. This answer sheets were used to know students' answer distribution. They were analyzed in order to find out their

validity, reliability, level of difficulty, and discrimination power to answer the statements problem no 1.

3.3.5 Data and Source of Data

In this study the data were obtained from the items of UAS test, the key answer, the students' answer sheets and curriculum of English for Fifth grade of Elementary School academic year 2011/2012.

3.4 Method of Collecting Data

To analyze the quantitative data to find its validity, reliability, discrimination power, and difficulty level, the writer collected the data from students' answer distribution. It was collected by recapitulating students' answers. It was done by writing down score 1 for correct answer and 0 for wrong answer. This method was used in multiple-choice and short-answer items. In essay-test, there was highest and lowest score for various type of answer. This scoring was done by standardizing the students' answer with the key answer.

Qualitative data was collected by observing the test-items of both test-packs. On the characteristics of a good test analysis, the data were taken from those items to see whether they fulfill the characteristics of a good test or not. They were separated by those which had been required the characteristics.

3.5 Method of Analyzing Data

3.5.1 Quantitative Data Analysis

Quantitative data analysis was done by analyzing students' answer manually. It was conducted by applying a formula of each item analysis (presented in paragraphs below) in *Microsoft Excel* for manual calculation. While, distribution of distractors could be calculated automatically using *ITEMAN* program.

3.5.1.1 Measuring the Validity

To know the validity of each number of the test-items, the writer used formula *product moment* as described below:

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

(Bachman, 2004:86 Tuckman, 1978: 163,)

Where:

- r_{xy} = correlation coefficient between variable X and Y
- N = number of test-takers
- $\sum X$ = number of test items
- $\sum Y$ = total score of test items
- $\sum XY$ = multiplication of items score and total score
- $\sum X^2$ = quadrate of number of test items
- $\sum Y^2$ = quadrate of total score of test items

3.5.1.2. Measuring the Reliability

To measure reliability of essay test items, the writer used the Alpha formula below:

$$r_{11} = \left(\frac{n}{n-1} \right) \left(1 - \frac{\sum \sigma_1^2}{\sigma_1^2} \right)$$

(Tuckman, 1978:163)

Where:

r_{11} : test of reliability

$\sum \sigma_1^2$: number of varians of each item test

σ_1^2 : test items' varians

N : total of test items

Classification of items reliability are:

$0,00 < r_{11} \leq 0,20$: very low

$0,20 < r_{11} \leq 0,40$: low

$0,40 < r_{11} \leq 0,60$: medium

$0,60 < r_{11} \leq 0,70$: high

$0,70 < r_{11} \leq 1$: very high

3.5.1.3 Measuring Level of Difficulty

Number that shows difficulty or easiness of a test items is known as difficulty index or level of difficulty. The formula used to measure it is:

$$IF = \frac{B}{JS}$$

(Brown, 2004:59)

Where:

IF = Item Facility (Level of difficulty)

B = number of test-takers answering the item incorrectly

JS = number of test-takers responding to that item

$$P = \frac{R}{N} \times 100$$

(Gronlund, 1993: 103 and Garrett, 1981:363)

Where

P = the percentage of examinees who answered items correctly.

R = the number of examinees who answered items correctly.

N = total number of examinees who tried the items.

When we are going to analyze the level of difficulty on essay tests or short answer item test which have the criteria of maximum and minimum score, different formula was used. The writer used the formula as stated below:

$$P = \frac{\text{Mean}}{\text{max imumscore}}$$

(Zulaiha, 2008: 34)

P : Level of difficulty of Essay test or Short Answer test

Mean : Average of students' score

Maximum Score : The maximum score of each item

Classifications of level difficulty of are:

$P = 0,00$: test items is too difficult

$0,00 < P \leq 0,30$: test items is difficult

$0,30 < P \leq 0,70$: test items is medium

$0,70 < P \leq 1,00$: test items is easy

$P = 1$: test items is too easy

3.5.1.4 Measuring Discrimination Power

There is no absolute P value that must be met to determine if an item should be included in the test as is, modified, or thrown out, but appropriate test item will generally have P that range between 0.15 and 0.85. (Brown, 2004:59)

The formula that can be used to measure the discrimination power of multiple-choice test items is:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} \quad (\text{Brown, 2004:59})$$

Where:

ID = Item Discrimination (Discrimination Power)

BA = number of top test takers that have correct answer

BB = number of bottom test takers that have correct answer

JA = total participant of top test-takers

JB = total participant of bottom test takers

The formula for computing item discrimination stated by Gronlund, (1993: 103) and (Ebel and Frisbie, 1991:231) is stated below:

$$D = \frac{R_v}{N_v} - \frac{R_L}{N_L}$$

Where

D = Index of discrimination.

RU= Number of examinees giving correct answers in the upper group.

RL = Number of examinees giving correct answers in the lower group.

NU or NL= Number of examinees in the upper or lower group respectively.

In measuring discrimination power of essay test, there was different formula used. It is different because each item of tests has highest and lowest score. To measure this, the writer used formula below:

(Zulaiha, 2008: 34)

$$D = \frac{MeanA - MeanB}{\max imumscore}$$

Where:

D : Discrimination Power

Mean A : the average of students' score on top group

Mean B : the average if students' score on bottom group

Maximum Score : the maximum score of each item

Classifications of test Discrimination Power are:

D = 0, 00 – 0, 20: poor Discrimination Power

$D = 0, 20 - 0, 40$: sufficient Discrimination Power

$D = 0, 40 - 0, 70$: good Discrimination Power

$D = 0, 70 - 1, 00$: very good Discrimination Power

$D = \text{negative}$, all of test items is not good. Thus, the items that have same negative

D score should be skipped.

3.5.1.5 Measuring Distractors' Distribution

The distribution of distractors means the distribution of alternative answers. The importance of calculating it is to know the students' answers. A good distractor is that it has the distribution index of more than 0.025 or 2,5%. If the index of this is 0, thus the distractor should be discarded. It can be found out by calculating manually or by using *ITEMAN* program. *ITEMAN* (Item and Test Analysis Manual) is a program that calculates a test with the output of several numbers of levels of difficulty, discriminating power, and distribution of distractors, reliability, failure measurement, and score distribution. Yet, this application can only be used in multiple-choice-question test type. Essay tests and short answer items cannot be analyzed by using this program. Thus, it does not need a formula to be applied in analyzing test as *Microsoft Excel* does. In this study, the first part of test-packs in which its type is multiple choice questions was analyzed by using this program. All of the quantitative aspects of MCQ in both test-packs would be all covered and found by entering students answer to it.

3.5.2 Qualitative analysis

It dealt with analysis and studied on non-statistical features on test items. There were two aspects on this sub chapter that the writer was going to study. They were analysis of instructional materials, and analysis of characteristics of a good test including analysis of language use in it.

3.5.2.1 Analysis of the Appropriateness to Curriculum

Analysis of instructional materials dealt with the appropriateness of the test items with instructional materials of teaching and learning process stated in curriculum as Standard and Basic competence. In this sub chapter, the test items were reviewed whether or not they have matched with Standard and Basic Competence especially on elementary school. In order to make the steps clear, the writer presented the illustration of the steps as follows:

Table 3.1: Curriculum Checklist

Competence Standard	Basic Competence	Indicator	Themes / Materials		Items test that appropriate with the basic competencies		Total number of items test (Σ)		Percentage of total numbers of particular items represent the elated basic competence	
			S 1	S 2	TP 1	TP 2	TP 1	TP 2	TP 1	TP 2
Listening 1. Students are able to understand very simple instruction with an action in school	1.1 Students are able to respond very simple instruction with logical action in	<ul style="list-style-type: none"> Students are able to complete a sentence in form of 								

Competence Standard	Basic Competence	Indicator	Themes / Materials		Items test that appropriate with the basic competencies		Total number of items test (Σ)		Percentage of total numbers of particular items represent the elated basic competence	
			S 1	S 2	TP 1	TP 2	TP 1	TP 2	TP 1	TP 2
context.	class and school context	Present Continuous Tense.								

In the table above, there are seven columns. The first and second columns contain Competence Standard and Basic Competence. On the next column, it fills with indicators of each Basic Competence. The themes and materials of the lesson are stated in the fourth column. In the fifth column, it contains the items of both test-packs that match to the Basic Competence. The total items that match to Basic Competence are stated in the sixth column. In the last column, it is filled by the percentage of the items match to Basic Competence.

3.5.2.2 *Analysis of the characteristics of a good test*

In analyzing of the characteristic of a good test, the writers used the characteristic of a good test check list, which contains some requirements to be said as a good test. It can be seen in the table below

**Table 3.2: Checklist of the Observation of the Characteristics of a
Good Test**

No	The characteristics of a good MCQ test	1	2	3	4	5	6	7
1	There should be one correct answer							
2	Only one feature at the time should be tested							
3	Each option should be grammatically correct when placed in stems							
4	Be efficient of using word, phrase, and sentences.							
5	Chronological on the optional answer							
6	It should not be dependent or other question							
7	The stems should not give clues or question to other question							
8	Be careful on capital letter							
9	All multiple choice items should be at a level appropriate to the proficiency level of education							
10	Pictures, Graphics, tables, and diagrams should be clear and in function							
11	The questions or statements should be grammatically correct							

Within this checklist, the writer observed each item of both test-packs to be fitted with the requirements of a good test in the table above. Then, if there was an item that did not have one or more requirements of a good test, it was analyzed on its error and the writer suggested for its improvement.

BIBLIOGRAPHY

- Arikunto, S. 2006. *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.
- Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. London: Oxford University Press.
- . 2004. *Statistical Analyses for Language Assessment*. London: Cambridge University Press.
- Bajracharya, I.K. 2010. "Selection Practice of the Basic Item for Achievement Test". In *Mathematics Education Forum*. Vol. 14 Issue 27 p 10-14.
- Best, John W. 1977. *Research in Education*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- Blood, D.F. and W.C. Budd. 1972. *Educational Measurement and Evaluation*. New York: Harper and Row.
- Brown, H. Douglas. 2002. *Principles of Language Learning and Teaching (4th Ed)*. New York: Addison Wesley Longman Inc.
- . 2004. *Language Assessment Principles and Classroom Practices*. San Francisco : Longman, Inc.
- BSNP. 2006. *Standar Isi dan Standar Kompetensi Lulusan Tingkat Sekolah Menengah Pertama dan Madrasah Tsanawiyah*. Jakarta: PT. Binatama Raya
- Celce-Murcia, Marianne and Elite Olshtain. 2000. *Discourse and Context in Language Teaching. A Guide for Language Teachers*. London: Cambridge University Press.
- Davies, Alan. 1997. "Demands of Being Professional in Language Testing." In *Language Testing 14th*. p 328-339.
- Direktorat Profesi Pendidik. 2008. *Standar Pengembangan Kelompok Kerja Guru (KKG) Musyawarah Guru Mata Pelajaran (MGMP)*. Jakarta: Departemen Pendidikan Nasional.
- Ebel, R. L. and Frisbie, D. A. 1991. *Essentials of Educational Measurement (5th ed.)*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

- Gronlund, N. E. 1993. *How to Make Achievement Tests and Assessments*. Boston: Allyn and Bacon.
- , 1998. *Assessment of Students Achievement*. 6th Edition. Boston: Allyn and Bacon.
- Hatch, E. and Farhady, H, 1982. *Research Design and Statistics for Applied Linguistics*. London: Newbury House Publishers, Inc.
- Hughes, A. 2005. *Testing for Language Teachers*. 2nd Ed. London: Cambridge University Press.
- Marshall, J. C. and Hales, L. W. 1972. *Essentials of Testing*. Massachusetts: Addison- Wesley Publishing Company Ltd
- Mehrens, W and Lehmen, I.J. 1984. *Measurement and Evaluation in Educational and Psychology*. New York: Halt Rinehart and Winston.
- Meizaliana. 2009. Teaching Structure through Games to the Students of Madrasah Aliyah Negeri I Kapahiang Bengkulu. M.Hum thesis. Diponegoro University.
- Muslich, Masnur. 2008. *KTSP (Kurikulum Tingkat Satuan Pendidikan) Dasar Pemahaman dan Pengembangan*. Jakarta: Bumi Aksara.
- , 2009. *KTSP Pembelajaran Berbasis Kompetensi dan Kontekstual*. Jakarta: Bumi Aksara.
- Nitko, A. J. 1983. *Educational Test and Measurement An Introduction*. New York: Harcourt Brace Jovanovich Publishers
- Nurulia, Lily. 2011. An Analysis of Multiple-choice English Formative Test for Grade VIII of MTsN 1 and MTsN 2 Semarang. M.Pd thesis. Semarang State University.
- Rammers H. H., Gage, N. L. and Rummel, J.I. 1967. *A Practical Introduction To Measurement and Evaluation* (2nd ed.). New Delhi: Universal Book Stall.
- Richards, Jack C. 2001. *Curriculum Development in Language Teaching*. London: Cambridge University Press.
- Tuckman, B. W. 1975. *Measuring Educational Outcomes Fundamentals of Testing*. New York: Harcourt Brace Javanovich Inc.

Valette, R.M. 1967. *Modern Language Testing*. 2nd Ed. New York: Harcourt Brace Jovanovich Publishers, Inc.

Zulaiha, Rahmah. 2008. *Analisis Soal Secara Manual*. Departemen Pendidikan Nasional Badan Penelitian dan Pengembangan Pusat Penilaian Pendidikan. Jakarta: PUSPENDIK.

APPENDICES