

**AN ANALYSIS OF CONTENT VALIDITY IN ENGLISH SEMESTER TEST  
ITEMS OF SENIOR HIGH SCHOOLS IN PADANG**

**THESIS**

**Submitted in Partial Fulfillment of the Requirement for Strata One (S1) Degree**



**By**

**Feni Sjafrianti  
64055/2005**

**Advisors:**

- 1. Prof. Dr. Mukhaiyar, M.Pd**
- 2. Dr. Kusni, M.Pd**

**ENGLISH DEPARTMENT  
FACULTY OF LANGUAGES AND ARTS  
STATE UNIVERSITY OF PADANG  
2011**

## **HALAMAN PERSETUJUAN**

**Judul** : An Analysis of Content Validity in English Semester Test  
Items of Senior High Schools in Padang

**Nama** : Feni Sjafrianti

**NIM** : 64055/2005

**Program Studi** : Pendidikan Bahasa Inggris

**Jurusan** : Bahasa dan Sastra Inggris

**Fakultas** : Bahasa Sastra dan Seni

**Padang, 7 Februari 2011**

**Disetujui oleh:**

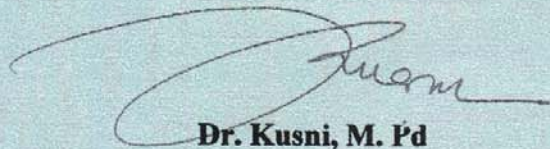
**Pembimbing I**



**Prof. Dr. Mukhaiyar, M.Pd**

**NIP: 19500612 197603 1 005**

**Pembimbing II**

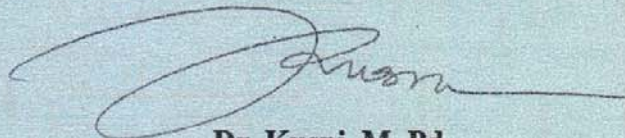


**Dr. Kusni, M. Pd**

**NIP: 19620909 198803 1 004**

**Diketahui**

**Ketua Jurusan Bahasa dan Sastra Inggris**



**Dr. Kusni, M. Pd**

**NIP: 19620909 198803 1 004**



## **HALAMAN PENGESAHAN LULUS UJIAN SKRIPSI**

**Dinyatakan Lulus Setelah Dipertahankan di Depan Tim Penguji Skripsi**

**Jurusan Bahasa dan Sastra Inggris**

**Fakultas Bahasa dan Seni**

**Universitas Negeri Padang**

**AN ANALYSIS OF CONTENT VALIDITY IN ENGLISH SEMESTER TEST**

**ITEMS OF SENIOR HIGH SCHOOLS IN PADANG**

**Nama : Feni Sjafrianti**  
**NIM/BP : 64055/2005**  
**Program Studi : Pendidikan Bahasa Inggris**  
**Jurusan : Bahasa dan Sastra Inggris**  
**Fakultas : Bahasa dan Seni**

**Padang, 7 Februari 2011**

**Tim Penguji**

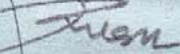
**Nama**

**Tanda Tangan**

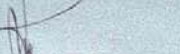
**1. Ketua : Prof. Dr. Mukhaiyar, M.Pd**

(  )

**2. Sekretaris: Dr. Kusni, M.Pd**

(  )

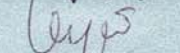
**3. Anggota : Dra. Yenni Rozimela, M.Ed, Ph.D**

(  )

**4. Anggota : Refnaldi, S.Pd, M.Litt**

(  )

**5. Anggota : Dra. Aryuliva Adnan, M.Pd**

(  )

## ABSTRAK

Sjafrianti, Feni. 2011. *An Analysis of Content Validity in English Semester Test Items of Senior High Schools in Padang*, Tesis.

Pembimbing : Prof. Dr. Mukhaiyar, M.Pd  
Dr. Kusni, M.Pd

Penelitian ini bertujuan untuk mengevaluasi tes semester bahasa Inggris kelas X Sekolah Menengah Atas (SMA) kota Padang. Hal ini dilakukan untuk menemukan dan menentukan validitas isi (*content validity*) dari tes melalui *content representativeness* atau keterwakilan *objectives* (indicator) dalam silabus ke dalam item-item (soal-soal) tes.

Jenis penelitian ini adalah penelitian evaluasi karena mencoba untuk mengevaluasi kualitas produk pendidikan yaitu tes dengan menggunakan analisa dokumen. Berdasarkan data, ini adalah penelitian kualitatif. Data berbentuk deskripsi terhadap objek-objek yang dievaluasi. Untuk mengevaluasi tes tersebut, peneliti menggunakan lembar pencocokan item-item tes sebagai instrumen.

Berdasarkan hasil penelitian, jelas bahwa tes tersebut memiliki *content validity* yang rendah. Dari 40 soal, hanya 14 soal yang cocok dengan indikator dalam silabus. Dengan kata lain, 26 soal tidak sesuai dengan indikator. Selain itu, dari 46 indikator yang terdapat dalam silabus, 34 diantaranya tidak direpresentasikan (tidak diukur) oleh soal-soal tersebut. Dalam sesi *listening*, terdapat enam soal dari 15 soal yang valid karena sesuai dengan indikator. Namun, terdapat 19 indikator dari 24 indikator *listening* yang tidak diuji dalam tes. Sesi berikutnya adalah *reading*. Dari 20 soal, hanya tujuh soal yang sesuai dengan indikator. Dari ketujuhanya, hanya empat indikator dalam silabus yang diuji. Padahal, terdapat 11 indikator dalam *reading*. Sesi *writing*, yang berbentuk essay, memiliki *content representativeness* yang paling rendah, sebab hanya satu soal (nomor 40) dari lima soal yang sesuai dengan indikator. 10 indikator lainnya tidak sesuai dengan soal-soal dalam sesi ini. Hasil ini menunjukkan bahwa item-item tes dari sesi *listening*, *reading* dan *writing* yang tidak mewakili *objectives*.

## **ACKNOWLEDGEMENTS**

The first and foremost, I offer my hearty devotion to Allah SWT, the Compassionate and the Merciful, for blessing that enabled me to finish this thesis. Besides, I also express my gratitude to the Prophet Muhammad SAW who led the human race to leave Jahiliyah era to the Islam era, which came together with religiosity, humanity and knowledge.

It is indeed a matter of great pleasure and satisfaction to record the constant guidance, encouragement and assistance received from a variety of sources from the conceptualization of the topic to the finalization of the research work.

At the first instance, I would like to express my deep sense of gratitude to Dr. Kusni, M.Pd and Prof. Dr. Mukhaiyar, M.Pd, my advisors for their kindness and patience, painstaking guidance, valuable suggestions and inspiration with intellectual stimuli to complete the research work. Then, I also thank to the examiners, Dra. Hj. Yenni Rozimela, M.Ed, Ph.D, Refnaldi, S.Pd, M.Litt, and Dra. Aryuliva Adnan, M.Pd, for the suggestions and critical review.

I would like to thank the chairperson of English Department, Dr. Kusni, M.Pd, the secretary of English Department, Dra. An Fauzia R. Syafei, M.A, and English Department administration staffs, and also my academic advisor, Yuli Tiarina, M. Pd for the supports and advices during my study. I am also grateful to all lectures of the English Department of State University of Padang who have guided and educated me patiently.

I am grateful to the head of MKKS Padang, Yunisra, M.Pd and the head of English Teacher Forum Padang, Seprah, M.Pd, who have given a very good cooperation. I also wish to express my gratitude to Neneng, S.Pd, for her kindness in helping me to collect the test sheets and tenth-grade level syllabus. Then, my gracious thanks go to my beloved friend and family for the spirit, suggestion and support.

Finally to my parents for their love, encouragement, understanding and patience during my long journey. I am forever indebted.

The writer,

Feni Sjafrianti



## TABLE OF CONTENTS

**HALAMAN PERSETUJUAN**

**HALAMAN PENGESAHAN**

**ABSTRACT**

**ABSTRAK**

**ACKNOWLEDGMENTS**

<b>TABLE OF CONTENTS .....</b>	<b>i</b>
<b>LIST OF TABLES .....</b>	<b>iii</b>
<b>LIST OF APPENDICES .....</b>	<b>iv</b>

### **CHAPTER I. INTRODUCTION**

A. Background of the Problem .....	1
B. Identification of the Problem .....	5
C. Limitation of the Problem .....	7
D. Formulation of the Problem .....	7
E. Research Questions .....	7
F. Purpose of the Research .....	8
G. The Significance of the Research .....	8
H. Definition of Key Terms .....	9

### **CHAPTER II. REVIEW OF THE RELATED LITERATURE**

A. Language Testing .....	10
B. Content Validity .....	21
C. Syllabus Design .....	24
D. Instructional Objectives.....	29
E. Review of Related Findings .....	31
F. Conceptual Framework .....	31

### **CHAPTER III RESEARCH METHODOLOGY**

A. Type of Research .....	33
---------------------------	----

B. Data and Source of Data .....	33
C. Research Instrument.....	34
D. Technique of Data Collection .....	34
E. Technique of Data Analysis .....	35
 <b>CHAPTER IV RESEARCH FINDINGS AND DISCUSSION</b>	
A. Findings .....	36
B. Discussion .....	50
 <b>CHAPTER V CONCLUSION AND SUGGESTION</b>	
A. Conclusions .....	56
B. Suggestions .....	57
 <b>BIBLIOGRAPHY.....</b>	<b>59</b>
<b>APPENDICES.....</b>	<b>61</b>



## **LIST OF TABLES**

Table 1. Items Match to Indicators of Listening Section .....	37
Table 2. Items Match to Indicators of Reading Section .....	40
Table 3. Items Match to Indicators of Writing Section .....	42
Table 4. Indicators Represented by the Test Items .....	43
Table 5. Indicators Unrepresented by the Test Items .....	44

## **LIST OF APPENDICES**

Appendix 1. The Test Items and Description of Its Objectives (indicators) based on the Test Items.....	61
Appendix 2. Check list on the Items Match to Indicators of Listening Section....	66
Appendix 3. Check list on the Items Match to Indicators of Reading Section ....	69
Appendix 4. Check list on the Items Match to Indicators of Writing Section.....	71
Appendix 5. English first semester test of Senior High Schools in Padang for tenth-grade level that designed by Teachers Forum (MGMP) of English Padang at academic year 2010/2011 sheet.....	72

# **CHAPTER I**

## **INTRODUCTION**

### **A. Background of the Problem**

Good teachers are consistently seeking better ways to present material and more efficient ways for the students to learn. In order to do this, they need to know something about their students' abilities, their past achievements, their interests, their strengths and their weaknesses. Evaluation, then, is valuable to the classroom teachers because it helps them to get information about their students.

A test is one kind of instruments designed for evaluation. It can provide much information about the general ability levels of students, about possible ability groupings, about specific problems that the students may have with the language, and about students' achievement in previous programs. It is constructed from a number of items (questions or problems). A test may contain all items of one type or a combination of item types. For example, the first type is the multiple-choice questions and the second type is essay questions.

Furthermore, a test should possess several qualities. Among the most important of these are reliability, validity and practicality. Practicality refers to the relationship between the resources that will be required in the design, development, and use of the test and the resource that will be available for these activities. A test can be called practical if a test is implemented within the means of financial limitations, time constraints, ease administration, scoring and interpretation.

Reliability concerns with the consistency of the test scores. Test can be called reliable if a same test is used to measure the students' achievement implemented in same students and in different situation, the result of the test is consistent across time.

By far the most complex criterion of a good test is validity. A test is said to be valid to the extent that it measures what it is supposed to measure. When a test does not accurately reflect what teachers have tried to teach, and when other factors, such as writing ability, are allowed to influence the student's test score, then the test is not valid.

Moreover, it includes daily test, mid-semester test, semester test, and final examination. The semester test of Senior High Schools in Padang is prepared and decided in the workshop of all headmasters of Senior High Schools (Musyawarah Kerja Kepala Sekolah/MKKS) Padang. There would be a decision here about who would design the test, how many items, day of test, duration of test, etc. This test measures students' achievement at the end of a semester in one academic year. The test is in form of multiple-choice and essay items.

This test is used only in all Senior High Schools in Padang. Thus, it can be categorized as a local test. The test designers are the English teachers from some schools who are selected or proposed by the headmaster from their school, or some English teachers who are selected from Teachers Forum (Musyawarah Guru Mata Pelajaran/MGMP) of English. The selected teachers are grouped in a team that consists of two or three teachers. Before designing the test, there are some requirements that the teachers need to consider in conducting the test. Those are



determining the level of students, making the table of test specification and peer analysis of the test. The time given to design a test is about two months. After finishing it, the test is submitted in compact disc to the MKKS Padang. Then, the MKKS will print and distribute it to all Senior High Schools in Padang.

After an interview with the head of English Teacher Forum Padang, the researcher got some information that there are some criteria in selecting teachers as the semester test constructors. The teachers should have good understanding of the subject matter on which test is to be made, sufficient knowledge of the students, adequate knowledge of different test formats that could be used, and also good knowledge about the characteristics of a good test, such as validity and reliability.

However, based on researcher observation on some last English semester tests from academic years 2008/2009 until 2009/2010 compiled by English Teacher Forum Padang, some problems occurred such as wrong key answers, misspelling, and ambiguous option. Moreover, since it is easy to develop items that require only recognition or recall of information in multiple-choice questions, they tend to rely heavily on this type of question. Whereas, the validity of multiple choice tests depends upon systematic selection of items with regard to both content and level of learning. Although most of the teachers try to select items that sample the range of content covered by the exam, they often fail to consider the level of the questions they select. From these facts, it can be seen that the last semester tests were not good enough.

Probably, this happened because of some reasons. First, the selected teachers are assumed to have already known about what a good test is, but in fact they have limited knowledge on it. Next, the teachers are not familiar with item analysis, such as difficulty index, discrimination index, function of the distractors, as procedures to make the test working well. Another reason is as what have been said by the head of MGMP in an interview with the researcher that the demands on teachers' time make them design the test without any review and critiques on the test items. This is in contrast with Hughes's statement (2003:3) that there should be a great deal of time and effort in constructing good tests. He also states that too many tests are written without care and attention. Then, the result is a set of poor items that cannot possibly provide accurate measurements.

Therefore, an analysis of test data can help evaluate and improve tests. As Hughes (2003:218) argues that analysis will provide the tester with useful information that may be used in making decisions about tests and test results. Then, the researcher wants to analyze one of the English semester tests of Senior High Schools in Padang and find out whether it needs improvement or not. Based on the analysis, we can also learn about how the quality of English test is.

The reason of choosing the tenth-grade level is related with the fundamental uses of achievements test in an educational program. It provides information for making decision whether the students can continue to the next grade or not, and also to see how far the objective of learning English have been achieved. Therefore, since the first year of study, the students' qualities must be reflected by the evaluation of

learning processes and achievements which can be seen by their scores in tests. So, there will be no wrong decision and students who continue to the next class or grade level can be considered as well-quality students.

## **B. Identification of the Problem**

Validity has been identified as the most important principle of a test (Brown, 2004). A test is said to be valid to the extent that it measures what it is supposed to measure. In order to examine the validity of a test, it requires a validation process by which a test user presents evidence to support the decisions made on the basis of test scores. Those evidences are construct validity, criterion-related validity, and content validity. For achievement tests, content validity evidence is the most important because the job of an achievement test is to measure how well the content taught has been mastered.

Some problems in investigating content validity have been identified by language testers (e.g., Bachman, 2002). First, difficulties may arise in defining the domain in a situation where examinees come from diverse backgrounds and have widely ranging needs in language use. Furthermore, even when the domain can be well defined, selecting representative samples from that domain may be problematic (Bachman, 2002). As pointed out by Hughes (1981), it is quite difficult to sample representative language skills as a result of inadequate needs analyses and the lack of comprehensive and complete descriptions of language use.

Moreover, Nitko (1996) explains some criteria that can be used when evaluating the test in relation to content representativeness and relevance. The first is whether the test items emphasize what have taught. This is in relation with the fact that the items on the test are often of poor quality. They emphasize low-level thinking skills, or emphasize different content than was emphasized during teaching.

The next criterion is whether test items accurately represent the outcomes specified in the school or national curriculum framework. Test that teachers used in grading should reflect the learning targets that the school and nation identify as important. Students' grades will be recorded and eventually be interpreted by people who have seen the curriculum, but who are not familiar with what the teachers taught in the classroom. They will expect the grades to reflect the national's learning targets. Since grades are based on teachers' tests, so the test items should reflect these learning outcomes.

The third is whether the test items are in line with the current thinking about what should be taught and how it should be assessed. Teachers, philosophers, curriculum theorists, researchers and others are constantly redefined what is worth learning. Professional teachers are aware with these developments and implement them in their teaching and testing practices.

The last criterion is whether the content in the test important and worth learning. However, the curriculum and content being taught contain many specifics. Teachers must be certain that the tested content relates directly to important student



learning targets. Content included in the test should also have great value or significance to a student's further learning or to a student's life skills.

### **C. Limitation of the Problem**

To make this research specific, the researcher only focuses on analyzing the content validity of the English semester test items of tenth-grade students of Senior High Schools in Padang at academic year 2010/2011 seen from its content representativeness. It focuses on whether the test items are a representative sample from a larger domain of performance.

The test contains of multiple-choice items and essays items. In analyzing this test, the researcher only focuses on the reading, listening and writing domains because the test only measure those skills.

### **D. Formulation of the Problem**

The problem is formulated in the following question: "What is the content validity of the English semester test items of the tenth-grade level of Senior High Schools in Padang at academic year 2010/2011 based on the judgments of content representativeness?"

### **E. Research Questions**

To make the formulation of the problem more specific, the questions above are developed into the following research questions:

1. Are the test items of listening section sample the objectives well?
2. Are the test items of reading section sample the objectives well?
3. Are the test items of writing section sample the objectives well?

#### **F. Purpose of the Research**

The purpose of this study is to evaluate English first semester test items of the tenth-grade level of Senior High Schools in Padang at academic year 2010/2011 based on the judgments of content representativeness. The research describes whether the test items of listening, reading and writing section sample the objectives well.

#### **G. Significance of the Research**

By presenting this research, the researcher hopes that there will be some benefits as follows:

##### **1. Theoretical Benefit**

This study gives contribution to the larger body of knowledge and additional information to teaching research especially those dealing with the content validity of test.

##### **2. Practical Benefit**

- a. This study can deepen the understanding in literary field as the reference and develop writer's skill and ability in conducting other researches.
- b. The finding of the research can help the teachers to organize, develop, or select the good test for their students.

## **H. Definition of Key Terms**

To avoid misunderstanding in this research, the key terms are defined as follows:

- 1. Test** : prepared administrative procedures that appear at identifiable times in a curriculum when learners muster all their abilities to show peak performance, knowing that their abilities are being evaluated and measured. (Brown, 2004)
- 2. Validity** : the criteria to measure a test whether it measures accurately what it is intended to measure. (Hughes, 2003)
- 3. Content Validity** : the degree to which a test's tasks and topical contents are relevant to, and proportionately representative of the real-life domain to which the test corresponds (Hughes, 1989; Bachman, 1990).
- 4. Content Representativeness** : concerns the extent to which the test accurately samples a larger domain of performance (Bachman, 2002).

## **CHAPTER II**

### **REVIEW OF RELATED LITERATURE**

#### **A. Language Testing**

##### **1. The Nature of Language Testing**

According to Bachman (1994), language testing is one of the tools to get valuable information about the effectiveness of learning and teaching. Language teachers regularly use tests to diagnose student strengths and weaknesses, to assess student progress and to assist in evaluating student achievement. Language tests are also frequently used as sources of information in evaluating the effectiveness of different approaches to language teaching. As sources of feedback on learning and teaching, language tests can thus provide useful input into the process of language teaching.

In addition, Bachman and Palmer (1997) state that it can also be used as a tool for clarifying instructional objectives and, in some cases, for evaluating the relevance of these objectives and the instructional materials and activities based on them to the language use needs of students following the program of instruction. For these reasons, virtually all language teaching programs involve some testing, and hence, language teachers need to be able either to make informed judgments in selecting appropriate language tests or to plan, construct, and develop appropriate tests of their own.



In conclusion, language test can be defined as an instrument for the measurement and evaluation of any knowledge, quality, or ability. It may measure degree or amount of achievement, mental abilities, personality and character traits. Moreover, language tests are used for a variety of purposes; these can be grouped into two broad categories. First, the results of language tests may be used to make inferences about test takers' language abilities or to make predictions about their capacity for using language to perform future tasks in contexts outside the test itself. Second, decisions (e.g., selection, diagnosis, placement, progress, grading, certification, employment) may be made about test takers on the basis of what can be inferred from test scores about their levels of ability or their capacity.

Test itself is a part of assessment. As Brown (2004) analyzes that it is only one among many procedures and tasks that many teachers can use to assess students. All tests are formal assessment. He says that they are systematic planned sampling techniques constructed to give teacher and student an appraisal of their achievement.

Furthermore, there are some competences in language testing as describes by Bachman and Palmer (1996). First, an understanding of the fundamental considerations that must be addressed at the start of any language testing effort, whether this involves the development of new test or the selection of existing language tests. Second, an understanding of the fundamental issues and concerns in the appropriate use of language test. Third is an understanding of the fundamental issues, approaches and methods used in measurement and evaluation. Fourth, the ability to design, develop, evaluate and use language tests in ways that are appropriate

for a given purpose, context and group of test takers. Last, the ability to critically read published research in language testing and information about published tests in order to make informed decisions.

According to Cohen (1994), conducting language test in the classroom is important to promote meaningful involvement of students with material that is central to the teaching objectives of a given course. Therefore, to achieve this meaningful involvement, the goals of the test tasks need to reflect the goals of the course, and these goals to be made clear to the students. He give such example that test may motivate students to pay closer attention to the material on a particular day, if the teacher announces at the outset of the class session that there will be a test on that material.

In order to make any decisions related to students' achievement and how to improve it, test constructors must have some ideas of the amount of language that each student is learning in a given period of time. Thus, tests design can be directly linked to the program goals and objectives. These achievement tests will typically administered at the end of the program to determine how effectively students have mastered the desired objectives.

## **2. Achievement Test**

The information gained in achievement test as states in Brown (1995) can be useful in reexamining the need analysis, in selecting or creating materials and teaching strategies, and in evaluating program effectiveness. Therefore, the

achievement test used must be very specific to the goals and objectives of a given program. Furthermore, the development of systematic achievement tests is very important to the evolution of a systematic curriculum.

Cohen (1980) states that an achievement test assesses what has been achieved or learned from what was taught in a particular course or a series of courses. Finocchiaro and Sako (1983) assume that achievement test is used to measure the amount and degree of control of discrete language and cultural items and of integrated language skills acquired by the student within a specific period of instruction in a specific course. Thus, achievement tests may be subdivided according to the time of the test administration and the scope of the material.

Brown (2004) implies that the primarily role of an achievement test is to determine whether the course objectives have been met by the end of a period of instruction. In addition, achievement tests are often summative because they are administered at the end of a unit or term of study.

Achievement tests can be used by teacher to provide information about the achievement of group of learners. Their aims will show how successful individual students, groups of students, and the courses have been in achieving objectives of specific study (Hughes, 2003). Moreover, achievement test scores are often used in an educational system to determine what level of instruction for which a student is prepared. High achievement scores usually indicate a mastery of grade-level material, and the readiness for advanced instruction. Low achievement scores can indicate the need for remediation or repeating a course grade.

Furthermore, Hughes (2003) states that achievement tests made by teachers can be divided in two types. First are final or semester achievement tests. It administered at the end of the course of study. Second, progress or mid-semester achievement tests which measures the progress that students are making. The content of both progress and achievements tests is generally based on the course syllabus or the course textbook. If the syllabus is badly designed, or the books and other materials are badly chosen, the results of a test can be very misleading. Successful performance on the test may not truly indicate successful achievement of course objectives.

When writing achievement test items, writers usually begin with a list of content standards (either written by content specialists or based on state-created content standards) which specify exactly what students are expected to learn in a given school year. The goal of item writers is to create test items that measure the most important skills and knowledge attained in a given grade-level. The number and type of test items written is determined by the grade-level of content standards. Content validity is determined by the representativeness of the items included on the final test.

### **3. Test Items**

A test is constructed from a number of items (questions or problems). The word *item* is used in preference to the word *question* because question suggests the interrogative form; whereas many test items, in fact, written in the form of statement.

Tenbrink (1974) states that there are five questions that a teacher should consider in developing a test item. First is about what types of items should be used. A test is constructed from a number of items (questions and problems). It may consist of all items of one type or a combination of item types. According to Tenbrink (1974), the type of test item to be used is determined by four factors. There are the level and kind of learning outcomes being measured, the way in which the results of the test will be used, the characteristics of the students taking the test and the time available for constructing, administering and scoring the test.

Furthermore, Tenbrink (1974) states that there are many ways to categorize test items. The most popular classification scheme categorizes test items along two dimensions, that is, according to the method of scoring and according to the freedom allowed in the students' response. Those tests which are scored objectively are called objective test items. The items that use more subjective scoring techniques are called subjective items, for instance is the essay test.

Then, according to the freedom that is allowed the student when making his response, test items can be classified into two, selection items and supply items. The selection item allows the student to select, from among some number of alternatives, for examples, true-false, multiple-choice and matching questions. The objective supply-type items are the fill-the-blank and short-answer questions. Since this research only focuses on multiple-choice item and essays items, so the rest of the test item types may not be discussed further.

### **a. Multiple-choice Items**

The most common kind of multiple-choice item present the examinee with a question along with four or five possible answers, from which one is to be selected. Brown (2004) explains that the initial part of multiple-choice item will typically be a question or an incomplete statement. It is known as the *stem*. Popham (1981) states that the stem could be a map, an illustration, graph or some other sort of presentation.

It is associated with either three, four or five answer options that only one of which is correct and several *distractors*. The function of the distractors as mentioned by Bailey (1998) is to distract the inattentive, unsure, or ill-prepared test taker, to lure him away from the key answer, or to provide plausible alternatives if he is guessing.

Furthermore, there are three things that should be considered in writing multiple-choice items as stated in Bailey (1998). The first is that although multiple-choice items can indeed be objectively scored, a great deal of subjective judgment goes into their development. Next, though they may be practical to score, good multiple-choice items are extremely labor-intensive to write. Finally, the most important thing is that there are also major concerns about the likely negative washback of using the multiple-choice format. The effect of testing on teaching and learning can be harmful. Hughes (2003) gives an example of it. If the skill of writing is tested only by multiple-choice items, there is great pressure to practice such items rather than practice the skill of writing itself.

Hughes (2003) says that perhaps the most obvious advantage of multiple-choice questions is that scoring can be perfectly reliable, objective and rapid and

economical as well. When test is carried out on a very large scale, when the scoring of tens of thousands of compositions might seem not to be a practical proposition, it is understandable that potentially greater accuracy is sacrificed for reasons of economy and convenience. Although the scoring can be objective, the designation of comprehension questions is not objective at all. The reliability is seldom high because there is the general disagreement on the determination of the most important part of the text on which most questions are based.

Hughes (1989) believes that another advantage is that “since in order to respond the candidate has only to make a mark in the paper, it is possible to include more items than would otherwise be possible in a given period of time”. His claim is theoretically applicable, but considering the time limit of the test, tests’ users cannot include as much items as they expected. It is possible to set the distractors so close that the pupil has to examine each alternative very carefully indeed before he can decide on the best answer. When a person answers a comprehension question incorrectly, the reason for his error may be intellectual or linguistic or a mixture of the two. Such errors can be analyzed and then classified so that questioning can take account of these areas of difficulty (Alderson, 2000). However, this positive attitude towards multiple-choice questions is problematic because answering multiple-choice items is an unreal task, as in real life one is rarely presented with four alternatives from which to make choice to signal understanding.

Alderson (1995) believes that it may be easier to control the thought process of readers with multiple-choice techniques than it is with short answer questions.

Because it is easier to devise distractors to get readers to think in certain ways and this control may be desirable for the testing of inference in the second language. But, this also implies that the method tricks the unwary into making incorrect interpretations they might not otherwise have made. It is just likely to test some abilities and not be so good at testing others.

Alderson (1995) agrees that the popular use of multiple-choice questions does not prove its validity. It has many disadvantages when the effects of this method are taken into account. It is evident that examinees taking multiple-choice tests can learn “strategies” that inflate their scores: techniques for guessing the correct answer, for eliminating implausible distractors, for avoiding two options that are similar in meaning, for selecting an option that is notably longer than the other distractors and so on. Hughes (1989) describes that practice at multiple-choice items (especially when, as happens, as much attention is paid to improving one’s educated guessing as to the content of the items) will not usually be the best way for students to improve their command of the language.

#### **b. Essay Items**

Popham (1981)) states that many teachers consider essay questions the ideal form of testing since essays seem to require more effort from the student than other types of questions. Students cannot answer an essay question correctly by simply recognizing the correct answer, nor can they study for an essay exam by memorizing



factual material. Essay questions can test complex thought processes, critical thinking, and problem solving.

Essay questions are best suited for testing the upper levels of cognition (analyzing, evaluating, creating), but these traits are unstable and often difficult to define. For example, is “critical thinking” the ability to construct a reasoned argument from evidence, to select the best course of action in a novel situation, to analyze weaknesses in competing arguments, or some combination of all of these things? If teachers wish to evaluate whether students have developed critical thinking skills in a course, the meaning of that phrase must be clearly defined, and the course objectives and essay test items should reflect the definition they have chosen.

Problem-solving skills can also be tested through essay items, but the format and method for solving problems must be specified by the teacher and clearly communicated to the student. Essay questions are often used in courses in which the development of writing skills is an important objective. But, again, one should establish the kinds of writing skills that students must demonstrate and provide some test time for thinking and for organizing the answer (otherwise, the combined effects of time pressure and test anxiety will usually result in poor writing). Of course, students should have sample opportunities to practice these skills before they have to demonstrate them on an exam.

According to Tenbrink (1974), the subjective essay items can also be classified along the dimension of freedom that allowed the student when making his response. They are a restricted-response essay item and an extended-response essay

item. The restricted-response essay allows the student to show how much information he can recall from memory. It is much more adapted to recall of facts, the listing of events which occur, or the recall of steps to be taken in a certain procedure. This kind of item can be constructed fairly easy in a short time. However, extended-response essay items are very difficult to score objectively. Well written of this item is not easy to write. Creative skill, ability to organize and present original ideas, or the ability to defend a position or to evaluate some product can be measured with this item.

The next question explained by Tenbrink (1981) is about how many items should be used. The number of items to be used is often determined by the amount of time available and particular situation. For example, if the students are slow readers, then one item per minute is too many. If the items are lengthy and require a great deal of thought, then this is also may be too many items per unit of time. A most important factor related to the length of a test is the reliability of a test. The longer the test, the higher the reliability is likely to be. Teachers should not only concern about the total length of the test, but also about the length of each part which measures the achievement of a different objective or level of achievement.

How the items should be presented is the next question. Objective test items are usually presented in booklet form (typed and duplicated). Essay items can be duplicated or in some cases written in the blackboard. Oral presentation is useful when the students are poor readers or have physical handicaps which make it difficult for them to read items from a booklet.

The last question is about how the students should respond. Objective tests can be answered more easily if students respond on separate answer sheets. Essay questions are usually responded to in writing, but an oral response can be valuable because a student can produce a much longer response in a short time. Of course, an oral response is more difficult to grade than a written one.

## **B. Content Validity**

Test validity is a critical factor to be judged in the total program of foreign language testing. Borg and Gall (1979) say that tests can be misused and may actually have deleterious effects on the person being tested if they do not have standards for validity. Validity is concerned with whether the information obtained from a test permits teacher to make correct decision about pupil's learning. Validation studies of language performance assessment are mainly concerned with three types of validity: construct validity, criterion-related validity, and content validity.

Content validity is a non-statistical type of validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured" (Anastasi & Urbina, 1997). A test has content validity built into it by careful selection of which items to include (Anastasi & Urbina, 1997). Items are chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. Foxcraft et al. (2004) note that by using a panel of experts to review the test

specifications and the selection of items the content validity of a test can be improved. The experts will be able to review the items and comment on whether the items cover a representative sample of the behaviour domain.

Moreover, content validity is of particular importance for achievement tests. A test score cannot accurately reflect a student's achievement if it does not measure what the student was taught and was supposed to learn. It can be compromised if the test covers topics not taught. Karmel and Karmel (1978) state, the teacher who gives an examination that covers the materials and objectives of instruction within her classroom has probably given a test that has content validity.

According to Hughes (2003), a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. The content must be relevant to the purpose of the measurement. This means, the items selected must be representative, sufficiently complete, uncontaminated with other content, and at the appropriate level of difficulty.

Cohen (1994) identifies that content validity is determined by checking the adequacy with which the test samples the content or objectives of the course or area being assessed. Sometimes, even commercial test developed by experts fail to state what objectives are being covered in the test and which items are specifically testing each of these objectives. For this reason, among others, teachers have been encouraged to look closely at what each test measures.

In short, Bachman (1990) says that content validity involves two crucial concepts: content relevance and content coverage. Content relevance refers to the extent to which the aspects of ability to be assessed are actually tested by the task, indicating the requirement to specify the ability domain and the test method facets. Content coverage concerns the extent to which the test tasks adequately demonstrate the performance in the target context, which may be achieved by randomly selecting representative samples. Bachman (2002) also assumes that the second aspect of content validity is similar to that of content representativeness, which also concerns the extent to which the test accurately samples the behavioral domain of interest.

However, Nitko (1996) explains some criteria that can be used when evaluating the test in relation to content representativeness and relevance. The first is whether the test items emphasize what have been taught. This is in relation with the fact that the items on the test are often of poor quality. They emphasize low-level thinking skills, or emphasize different content than was emphasized during teaching.

The next criterion is whether test items accurately represent the outcomes specified in the school or national curriculum framework. Test that teachers used in grading should reflect the learning targets that the school and nation identify as important. Students' grades will be recorded and eventually be interpreted by people who have seen the curriculum, but who are not familiar with what the teachers taught in the classroom. They will expect the grades to reflect the national's learning targets. Since grades are based on teachers' tests, so the test items should reflect these learning outcomes.

The third is whether the test items are in line with the current thinking about what should be taught and how it should be assessed. Teachers, philosophers, curriculum theorists, researchers and others are constantly redefined what is worth learning. Professional teachers are aware with these developments and implement them in their teaching and testing practices.

The last criterion is whether the content in the test important and worth learning. However, the curriculum and content being taught contain many specifics. Teachers must be certain that the tested content relates directly to important student learning targets. Content included in the test should also have great value or significance to a student's further learning or to a student's life skills.

Some problems in investigating content validity have been identified by language testers (e.g., Bachman, 2002). First, difficulties may arise in defining the domain in a situation where examinees come from diverse backgrounds and have widely ranging needs in language use. Furthermore, selecting representative samples from that domain may be problematic. As pointed out by Hughes (1981), it is quite difficult to sample representative language skills as a result of inadequate needs analyses and the lack of comprehensive and complete descriptions of language use.

### **C. Syllabus Design**

Syllabus is an important part of learning process. Indonesian Government Regulation No. 19 of 2005 on National Education Standards Article 20, says that "Planning the learning process includes the syllabus and learning implementation

plan (rencana pelaksanaan pembelajaran/RPP) that includes at least the purpose of learning, teaching materials, teaching methods, learning resources, and assessment of learning outcomes."

According to Committee of National Education Standards (Badan Standar Nasional Pendidikan/BSNP, 2006) syllabus is a lesson plan on a and / or groups of subjects/themes that include certain standard of competence, basic competence, subject matter/learning, learning activities, indicators for assessing the achievement of competence, assessment, allocation of time and learning resources. Syllabus answer questions about what kind of competencies that must be mastered by students, how to achieve it and how to find out achievement.

Brown (1995) defines syllabus design as selection and organization of instructional content including suggested strategy for presenting content and evaluation. Whereas, Curriculum is a broad description of general goals by indicating an overall educational-cultural philosophy which applies across subjects together with a theoretical orientation to language and language learning. However, syllabus is a detailed and operational statement of teaching and learning elements which translates the philosophy of the curriculum into a series of planned steps leading towards more narrowly defined objectives at each level. Moreover, Brown (1996) states that syllabus must be developed in order to address students' needs, to actualize the institutional goals and objectives, and to develop content standard (standard competencies and basic competencies into teachable materials used in teaching and learning process in related schools).

Indonesian Government Regulation No. 19 of 2005 on National Education Standards Article 17 paragraph (2), states that “Schools and school committees, or madrassas and madrassa committee, develop School Based curriculum and syllabus based on the basic framework of the curriculum and competency standards, under the supervision of district that is responsible in the field of education for elementary, junior high, high school, and SMK, and the department that handles government affairs in the field of religion for MI, MTs, MA, and MAK.” Thus, the components of syllabus consist of competency standards, basic competence, main material/learning, learning activities, indicator, assessment, time allocation, and learning resources. It can be developed by a group of teachers in one school, National Council of Teachers of English (MGMP), curriculum developer and other related resource persons, and also supervisor.

Committee of National Education Standards (Badan Standar Nasional Pendidikan/BSNP, 2006) states eight steps in syllabus development. First is assessing and determining competency standards. Assessing the competency standards subject should notice; (a) sequence based on concept hierarchy of disciplines and/or level of difficulty of material, not necessarily in the order that is in content standards (Standar Isi/SI); (b) linkages between standards of competence and basic competence in the subject; (c) relevance of standard of competence and basic competence between lesson.

Second step is assessing and defining basic competencies. This should consider the following points: (a) sequence based on concept hierarchy of disciplines



and/or level of difficulty of material, not necessarily in the order that exists in SI; (b) linkages between standards of competence and basic competence in the subject; (c) relevance standard of competence and basic competence between subjects.

Third step is identifying material/learning. Identify the subject matter should consider: (1) potential learners; (2) relevance to regional characteristics; (3) level of physical, intellectual, emotional, social, and spiritual learners; (4) usefulness for learners; (5) structure of science; (6) Timeliness, depth, and breadth of learning materials; (7) relevance to the needs of learners and demands of the environment; (8) time allocation.

Fourth step develop learning activities. Learning activities designed to provide learning experiences that involve mental and physical processes through interaction between learners, learners with teachers, environmental, and other learning resources in the achievement of competence. Learning experience intended to accomplish through a varied approach to learning and learner centered. Learning experience includes life skills that need learner should have.

Fifth step is formulating indicators of competencies achievement. The indicator is a marker of the achievement of basic competencies that are marked by changes in behavior that can be measured; include attitude, knowledge, and skills. Indicator developed in accordance with the characteristics of learners, educational units, and potential areas. It is used as a basis to develop assessment tools. In developing indicators, every basic competence developed into several indicators (more than two). Indicators use operational verbs that can be measured and/or

observed. Level of verb in the indicator is lower than or equivalent to the verb in basic competency and standard competency.

Sixth step is determining the type of assessment. Assessment is a series of activities to acquire, analyze, and interpret data about the process and learning outcomes of students who carried out systematically and continuously, so that it becomes meaningful information for decision making. The assessment was conducted by using tests and non-test in the form of written or oral, observation of performance, attitude, assessment of the work of a project or product, the use of portfolios, and self-assessment.

Seventh is determining the allocation of time. The allocation of time on each competency is based on the number of effective weeks and time allocation of subject per week by considering the number of basic competencies, breadth, depth, complexity, and the importance of basic competence. Allocation of time specified in the syllabus is the average time expected to master the basic competencies needed by a variety learners.

Eighth step is determining the source of learning. Learning source is a reference source, object and/or materials used for learning activities. Learning resources can be printed and electronic media, resource persons, as well as physical environment, natural, social, and cultural. Determination of learning resources based on standards of competence and basic competencies and subject matter/learning, learning activities, and indicators of competencies achievement.

#### **D. Instructional Objectives**

In Indonesian newest curriculum, that is, School Based Curriculum (Kurikulum Tingkat Satuan Pendidikan/KTSP), objective named as indicator. It is one of the factors affecting syllabus design and choice. Brown (1995) defines objectives as specific statements that describe the particular knowledge, behaviors and/or skills that the learner will be expected to know or perform at the end of a course or program.

Moreover, Richards (2001) defines objectives as a more precise focus to program goals. The objective consists of a statement of specific changes a program seeks to bring about and results from an analysis of the aim into its different components. Thus, objectives can also be seen as precise statements about what content or skills the students must master in order to attain a particular goal. In summary, objectives define what are to teach, and test tells the degree to which the goals are being realized.

Lindvall (1961) states that the translation of semester or course goals into the more specific and immediate objectives can guide the teacher in daily teaching. These statements provide teacher with clear and complete guidance on what is to be taught and also provide the key to the evaluation of achievement by indicating what the pupils are to be able to do when the learning is completed.

Test serves as a means of defining and evaluating the instructional objectives (they are also referred to as course objectives, learning objectives or teaching objectives). Valette (1977) implies that the classroom tests define the short-range

course objectives of the teacher in a very real way. Lindvall and Nitko (1975) also state that the information about the degree to which pupils have achieved the specified instructional objectives is desired in the evaluation of pupil achievement. Mager (1975) points out that if objectives are not clear, test are misleading, and at worst, they are irrelevant, unfair, or uninformative.

Objectives usable in test construction must be different from those with which many teachers have been familiar in the past. First, they must be stated in terms of observable student behavior. Therefore, the objective must be stated in terms of behavior that teacher will ask students to demonstrate. Next, the objective must be specific (Smith and Adams, 1966).

Richards (2001) states some general characteristics of objectives. First, they describe what the goal seeks to achieve in terms of smaller units of learning. Next, they provide a basis for the organization of teaching activities. Last, they describe learning in terms of observable behavior or performance.

Kemp (1977) implies that objectives for learning can be grouped into three major categories. They are cognitive, psychomotor and affective. The cognitive domain includes objectives concerning knowledge, information, thinking, naming, recognizing, predicting, etc. The psychomotor domain treats the skills requiring use and coordination of skeletal muscles, as in physical activities of performing, manipulating and constructing. The affective domain involves objectives concerning attitudes, appreciations, values, and all emotions such as enjoying, conserving, respecting, etc.

## **E. Review of Related Findings**

These relevant studies are used to elaborate the related research that had been done by other researchers as a reference for the researcher herself. Wartu (2007) conducted a research about test validity of teacher-made test. She found out that the test was invalid.

Another research was conducted by Subandi (2001). The aims of the research were to analyze degree of the difficulty, discriminating power, and distractors of the end semester test of Educational Profession Courses at State University of Padang. It was found that it needs elaboration for some discussed topics. It also suggested that lectures be responsible for the course redesign, and improvement of the concerned topics.

These researches are taken as references to see how other researchers had done the tests analysis. This research is conducted to analyzes content validity of English first semester test items for the tenth-grade students of Senior High Schools in Padang at academic year 2010/2011 sees from the content representativeness.

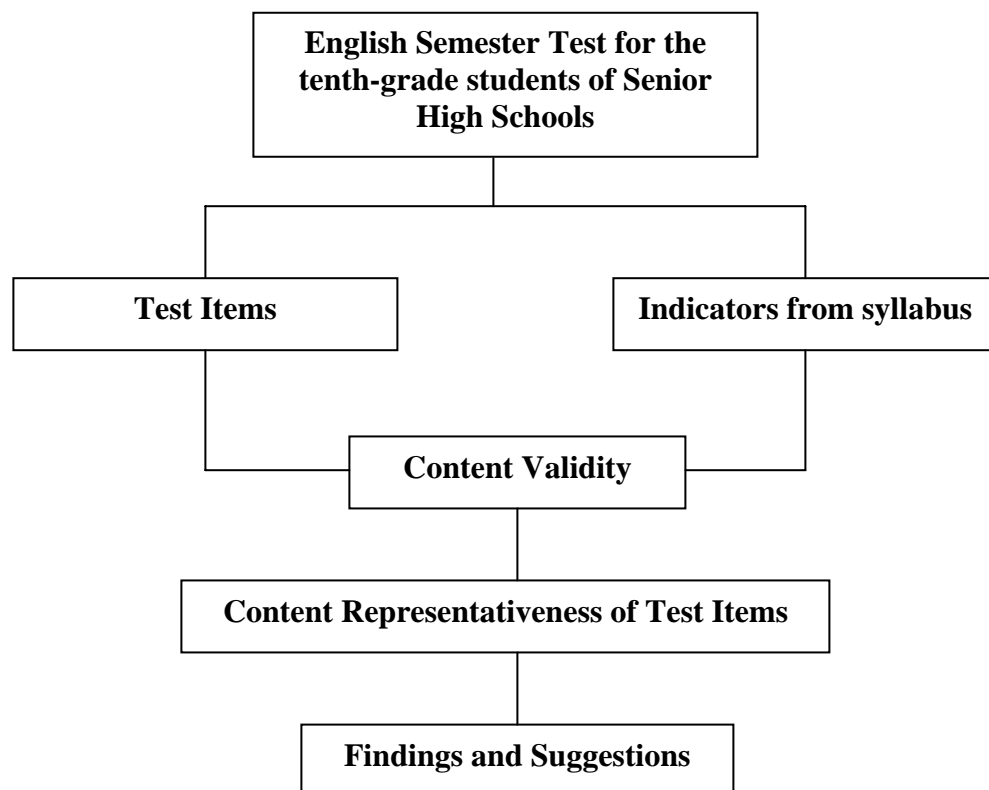
## **F. Conceptual Framework**

The researcher wants to analyze one of the English semester tests of Senior High Schools in Padang and find out whether it needs improvement or not. Validity has been identified as the most important principle of a test. One primary validity concept in order to ascertain how thoroughly a test items samples the instructional

objectives is content validity. The best way to ensure that a test has content validity evidence is to be sure its items match or measure the instructional objectives.

The research analyzes content validity of English first semester test items for the tenth-grade level of Senior High Schools in Padang at academic year 2010/2011 sees from its content representativeness. It is whether the test items emphasize what have taught. This is in relation with the fact that the items on the test are often of poor quality. They emphasize low-level thinking skills, or emphasize different content than was emphasized during teaching.

Understanding the review of related theories and findings discussed, the researcher formulates her theoretical framework as follows;



## **CHAPTER V**

### **CONCLUSION AND SUGGESTION**

#### **A. Conclusions**

Based on the result of this research, it can be concluded that the test items of English first semester test for the tenth-grade level of Senior High Schools in Padang at academic year 2010/2011 has low content validity. From 40 test items, only 14 items that match with objectives (indicators) in syllabus. In other words, 26 items do not match with the indicators. Moreover, from 46 indicators mentioned in syllabus, 34 of it are not represented (not measured) by the test items.

In listening section, there are six items from 15 items that are valid. It matches with the indicators. They are items number 1, 2, 3, 4, 5 and 15. There are three items (1, 3, 4) that ask students' to give response related to expressions in the dialogue that have heard. Items 2 and 5 identify meaning from expressions of happiness, sympathy and refuse appointment spoken in the dialogue. Only item number 15 that relates to text, that is, asks students to identify the material used in procedure text. However, 19 indicators from 24 listening indicators that are not measured by the test.

The next section is reading. From 20 items, only seven items that represent the indicators. From the seven items, only four indicators in syllabus that represent by the items. Yet, there are 11 indicators from syllabus that should be measured. seven items. Three items (number 16, 25, 36) ask about the same indicator. It is about identify the purpose of the text. The four other items (number 17, 19, 26, 34) ask to

identify topic of the text, main idea from a paragraph, order of event in text, and figure from story.

The writing section, that is in form of essay, considers as a very low content representativeness, because only one test item (number 40) from five items that match with the indicators. The others 10 indicators in the syllabus are not match with the test items.

These results then describe that the test items of listening, reading and writing section do not well-sampled the objectives.

## **B. Suggestions**

In accordance with the research findings, some suggestions are proposed for getting the better result in conducting an English semester test that contains a good content validity, as follows:

1. Teachers as test designers have to identify clearly the important learning targets and be sure that they are well-sampled by the test items. In conducting a semester test, teachers should make sure that the learning outcome specified in the test question must match the learning outcome described in the objective. This rule will ensure that the test teachers are developing will have content validity.
2. Teacher as test constructors have to ensure that the test items accurately represent the outcomes specified in the school or national curriculum framework. Test that teachers used in grading should reflect the learning



targets that the school and nation identify as important. Students' grades will be recorded and eventually be interpreted by people who have seen the curriculum, but who are not familiar with what the teachers taught in the classroom. They will expect the grades to reflect the national's learning targets. Since grades are based on teachers' tests, so the test items should reflect these learning outcomes. Another reason is because the implementation of school-based curriculum between one school to other schools is different because the development of it depends on the capability of its school, the potential and characteristics of region, and also the social cultural background of the students. So, the test items should be the representatives them.

3. Then, it is hoped that the school or English Teacher's Forum can set a training program for designing test, so the teacher has many chance to learn about designing a test.
4. In selecting teachers who will construct the test, it is better for the Teacher Forum (MGMP) of English to make sure that they have qualification on designing test, especially good knowledge about the characteristics of a good test, such as content validity.
5. Currently teachers' needs are the ability to criticize the tests. It means, never use the test without any review on the test items, because the result is a set of poor items that cannot possibly provide accurate measurements. Therefore, it is better if the teachers are provided by the sufficient time so they are able to review and criticize the construction or content of the test.

## Bibliography

- Bachman, Lyle F. and Adrian S. Palmer. 1996. *Language Testing in Practice*. London: Oxford University Press.
- Bailey, Kathleen. M. 1998. *Learning about Language Assessment*. New York: Heinle & Heinle Publishers.
- Borg, Walter R. and Meredith Damien Gall. 1979. *Educational Research: An Introduction*. New York: Longman, Inc.
- Brown, H. Douglas. 2004. *Language Assessment: Principles and Classroom Practices*. San Francisco: Pearson Education, Inc.
- Brown, James Dean. 1995. *The Elements of Language Curriculum: A Systematic Approach to Program Development*. Boston: Heinle & Heinle Publishers.
- Cohen, Andrew D. 1980. *Testing Language Ability in the Classroom*. Massachusetts: Newbury House Publishers, Inc.
- Cohen, Andrew D. 1994. *Assessing Language Ability in the Classroom*. Boston: Heinle & Heinle Publishers.
- Gay, L.R., and Peter Airasian. 2000. *Educational Research: Competencies for Analysis and Application*. New Jersey: Prentice-Hill, Inc.
- Finocchiaro, Mary and Sydney Sako. 1983. *Foreign Language Testing: A Practical Approach*. New York: Regents Publishing Company, Inc.
- Hughes, Arthur. 2003. *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Karmel, Louis J. and Marilyn O. Karmel. 1978. *Measurement and Evaluation in the School*. New York: MacMillan Publishing Co., Inc.
- Lindvall, C. M, and Anthony J. Nitko. 1975. *Measuring Pupil Achievement and Aptitude*. New York: Harcourt Brace Jovanovich, Inc.
- Popham, W. James. 1981. *Modern educational Measurement*. New Jersey: Prentice-Hill, Inc.

- Richards, J.C. 2001. *Curriculum Development in Language Teaching*. Cambridge: Cambridge University Press.
- Sabandi, Ahmad. 2001. *Analisis Butir Soal dan Pengembangan Mata Kuliah: Forum Pendidikan No.04 Tahun 26/ Edisi Desember 2001*. Padang: Universitas Negeri Padang Press.
- Smith, Fred M. 1966. *Educational Measurement for Classroom Teacher*. New York: Harper & Row Publishers.
- TenBrink, Terry D. 1974. *Evaluation: A Practical Guide for Teachers*. New York: McGraw-Hill, Inc.
- Warti. 2007. *Analisis Validitas Tes Bahasa Indonesia Semester Genap Kelas VIII SMP N 2 Lubuk Basung Tahun Pelajaran 2006/2007: Unpublished thesis*. Padang: Universitas Negeri Padang.