CHAPTER II

LITERATURE REVIEW

This chapter is an overview of the theory related to the study. It consists of evaluation, test, and validity. These points are discussed to give understanding about the topic being studied.

2.1 Evaluation

Evaluation is a part of a system that is useful for the enhancement of teachers teaching and students learning processes (Sulistyo, 2009). He added, in this way evaluation has corrective measures that will have backwash effects on the performance of the teachers and students alike. Teachers use assessments and evaluation to provide students and parents with ongoing feedback, plan further instructional and learning activities, set subsequent learning goals and identify students who may require intervention (British Columbia Ministry of Education, 2004)

Evaluation is viewed as 'a process of collecting information about different aspects of a language program in order to understand how the program works and how successfully it works, so different kinds of decisions can be made' (Richards, 2001). The main purpose of the evaluation is to guide classroom instruction and enhance student learning on a day-to-day basis (Genesee, 2001). Considering all aspects of learning and teaching, program evaluation might be either formative, developmental or on-going, which takes place during the course, and summative, which is conducted at end of the course

(Rea-Dickins & Germaine, 1992). While the formative evaluation is usually informal and aims to refine the existing program by making necessary adjustments, the summative evaluation is formal and its purpose is to assess how effective and efficient the program is. Summative evaluations that are periodically conducted provide useful information about what has been accomplished and "put a program in an excellent position to respond to crises when and if, they occur" (Brown, 1989).

Evaluation has two primary purposes: "accountability or summative evaluation; improvement or formative evaluation" (Torres, 2010). Formative evaluation 'provides information about how a program or organization operates and how to improve it' (Torres, 2010). This type of evaluation done for the purpose of improvement focusing on implementation and process and it is conducted while the program under study is ongoing or in the development stage (Mathison, 2010). Summative evaluation 'provides information about the overall effectiveness, impact, and/or outcomes of a program' (Torres, 2010). This type of evaluation serves accountability purposes and focuses on outcomes and effects, it is conducted when the program under study is completed or is in its final form (Mathison, 2010).

2.2 Test

According to Brown (1987) test is a method of measuring a person's ability or knowledge in a given area. The information provided by testing is essential to effective formal education and that this feedback conveys appropriate changes in the program that improve learning and teaching (Bachman, 1990)

2.2.1 Language Testing

"Language tests can be valuable sources of information about the effectiveness of learning and teaching. They can also be used as a source of feedback on learning and teaching." (Bachman, 1990)

McNamara (2000) pointed out the types of test differ with respect to how they are designed, and what they are for: in other words, in respect to test *method* and *purpose*. He (2000) stated that in terms of method, we can broadly distinguish traditional paper-and-pencil language tests from performance tests. Paper-and-pencil tests take the form of the familiar examination question paper. They are typically used for the assessment either of separate components of language knowledge (grammar, vocabulary, etc.) or of receptive understanding (listening and reading comprehension). In performance based test, language skills are assessed in act of communication. Performance tests are most commonly tests of speaking and writing, in which a more or less extended sample of speech or writing is elicited from the test-taker, and judged by one or more trained raters using an agreed rating-procedure. These samples are elicited in the context of simulations of real-world tasks in realistic contexts.

He (2000) added that tests' purpose itself differs between achievement and proficiency tests. Achievement tests are associated with the process of instruction such as end of course tests, portfolio assessments, or observational procedures for recording progress on the

basis of classroom work and participation. Achievement tests should support the teaching to which they relate. Whereas achievement tests relate to the past in that they measure what language the students have learned as a result of teaching, *proficiency tests* look to the future situation of language use without necessarily any reference to the previous process of teaching.

Achievement tests look backwards in that they assess what should already have been learnt, proficiency tests tend to look forward in that they assess a person's language skills and allow for interpretations of their future performance to be made (Kluitmann, 2008)

2.2.2 What should be tested in language?

Linguists are examining the whole complex system of language skills and patterns of linguistic behavior. Indeed, language skills are so complex and so closely related to the total context in which they are used as well as too many non-linguistic skills (gestures eye –movements, etc) that it may often seem impossible to separate them for the purpose of any kind of assessment (Heaton, 1988). He lists out the following ways of assessing performance in the four major skills:

> *Listening (auditory) comprehension in which short utterances dialogue talks and lectures are given to the testes, speaking ability usually in the form of an interview, a picture description, role play and a problem solving task involving pair work or group work, reading comprehension in which questions are set to test the students ability to understand the gist of a text and to extract key information on specific points in the text and writing ability usually in the form of letters reports memos messages- instructions and accounts of past events etc.*

He (1988) also notes that "---it is usually extremely difficult to separate one skill from another, for the very division of the four skill, is an artificial one and the concept itself constitutes a vast oversimplification of the issues involved in communication."

In agreement with this, Harrison (1989) stated, "Test of a foreign language should seek for useful information about functional language ability, general language proficiency and some areas of linguistic knowledge"

2.2.3 Standardize Test

George (2015) defined standardized test as any form of test that (1) requires all test takers to answer the same questions, or a selection of questions from common bank of questions, in the same way, and that (2) is scored in a "standard" or consistent manner, which makes it possible to compare the relative performance of individual students or groups of students. While different types of tests and assessments may be "standardized" in this way, the term is primarily associated with large-scale tests administered to large populations of students, such as a multiple-choice test given to all the eighth-grade public-school students in a particular state, for example.

In addition to the familiar multiple-choice format, standardized tests can include true-false questions, short-answer questions, essay questions, or a mix of question types. While

standardized tests were traditionally presented on paper and completed using pencils, and many still are, they are increasingly being administered on computers connected to online programs. While standardized tests may come in a variety of forms, multiple-choice and true-false formats are widely used for large-scale testing situations because computers can score them quickly, consistently, and inexpensively. In contrast, open-ended essay questions need to be scored by humans using a common set of guidelines or rubrics to promote consistent evaluations from essay to essay—a less efficient and more time-intensive and costly option that is also considered to be more subjective. One example of standardized test is national examination.

2.2.4 National Examination in Indonesia

A national assessment is designed to describe the achievement of students in a curriculum area collected to provide an estimate of the achievement level in the education system as a whole at a particular age or grade level. It provides data for a type of national education audit carried out to inform policy makers about key aspects of the system (Greaney & Kellaghan, 2008).

In Indonesia, high-school centralized tests have been administered since 1980. They were called EBTANAS (*Evaluasi Belajar Tahap Akhir Nasional* or National Final Evaluation of Students‟ Learning) from 1980 to 2001, and then UAN (*Ujian Akhir*

Nasional or National Final Examination) in 2002. They have later been named UN (*Ujian Nasional* = National Examination) since 2005 (Umam, 2011). Government Regulation of the Republic of Indonesia No. 19 about Education National Standard year 2005, Ministerial Regulation of National Education and Culture No. 66 about Educational Evaluation Standard year 2013, and Ministerial Regulation of National Education and Culture No. 3 about Students' Passing Criteria year 2013 define National Examination as an activity that measures the students' competence of certain subjects to evaluate the achievement of the National Education Standard that is nationally held every academic year. The passing decision made on the basis of National Examination scores is based upon a criterion-referenced decision. A criterion-referenced decision evaluates an examinee's performance based on a particular standard score on an examination that serves as a predetermined criterion (Brown, 2005).

National assessment systems in various parts of the world tend to have common features. All include an assessment of students' language or literacy and of students' mathematics abilities or numeracy. Some systems assess students' achievements in a second language, science, art, music, or social studies (Greaney & Kellaghan 2008).

As for Indonesia, described in Government Regulation of the Republic of Indonesia No. 32 about Education National Standard year 2013 and Ministerial Regulation of National Education and Culture No. 66 about Educational Evaluation Standard year 2013, national examination for senior high school in Indonesia aims to nationally measure students' competence on school subjects, namely Indonesian language, English language, Mathematics, and subjects which are specially characterized for the education program.

2.2.4.1 National English Examination 2015/2016 for senior high school

National English examination for senior high school in academic year 2015/2016 was held on April 4-6, 2016. The test's material was made based on criteria of graduation competence, content standard, and curriculum of 2013 and 2006 or KTSP (BSNP, 2016).

The material covers the level cognitive of knowledge, and understanding, application, and reasoning. In the level of knowledge and understanding, the students were asked to identify the topic/purpose/background/reason from short functional texts (announcement, letter, news, biography, procedure) and essay texts (recount, narrative, report, analytical exposition, news items, discussion). While in the level cognitive of application, the students were asked to classify, decide, and

apply the detail function of every step/tools/events/parts/aspects which were mentioned in the text given. And in the level cognitive of reasoning, the students were asked to conclude and analyze the text given (BSNP, 2016).

2.3 Validity

Gronlund (1990) emphasized validity as a matter of degree, it does not exist on an all – or none basis. Consequently, we should avoid thinking of evaluation results as valid or invalid. Validity is best considered in terms of categories that specify degree, such as high validity, moderate validity and low validity. Validity is always specific to some particular use or interpretation. No test is valid for all purposes. This is because evaluation results have a different degree of validity for each interpretation to be made.

In language testing, validating a test means being able to establish a reasonable link between a test-takers performance and their actual language ability. So, the question in validating a test is: "Does the test measure what it is intended to measure?" (Lado 1965).

Validity can be seen as a concept that allowing us to give test scores with meaning. This unitary notion of validity has traditionally been subdivided according to the kind of evidence on which the interpretations are based. Usually, one will come across the terms 'construct validity', 'content validity', 'criterion-oriented validity', 'concurrent validity', 'face validity' and 'consequential validity'. It should, however, be understood "that these 'types'

are in reality different ‘methods’ of assessing validity” and “that it is best to validate a test in as many ways as possible” (Alderson, et al. 2005).

2.3.1 Content Validity

Content validity is the degree to which a test's tasks and topical contents are relevant to, and proportionately representative the real-life of the test takers (Hughes, 1989). It is concerned with whether the content of a test is capable to gain information that is representative of the entire domains or skills, understandings, and other behavior that the test is supposed to measure (Aiken, 2000).

Bachman (1990) identified two aspects of content validity: content relevance and content coverage. In this case, content relevance does not only refer to the abilities the test aims to measure, but also to the test method, something which “is often ignored” (Bachman 1990). Content validity can be evaluated in part by showing the relevance of tasks and topical contents to the construct being tested and/or to the 'target language use' (TLU) domain, i.e. the real-life situation in which the language will be used (Bachman and Palmer, 1996).

Furthermore, Harrison (1989) pointed out that content validity is concerned with what goes into the test. He (1983) added that the content of a test should be decided by considering the purposes of the assessment, and then drawing up a list known as a content specification.

In agreement with Harrison, Abayomi (1999) asserts that the table of specifications or content specification ensures the content

validity of a test right from the construction stage. All these imply that for a test to be valid, the test planner should aim at a systematic coverage of the whole subject matter area and the instructional objectives.