CHAPTER II

LITERATURE REVIEW

2.1 Theoretical Framework

In this chapter, the writer would cover about the theory that used to describe the phenomena mentioned in previous chapter and also used to define the subject of the next chapter. In this chapter, it will divided into two parts, the first part will explains about the theories used, and the next part will cover about the comparison studies to support this thesis.

2.1.1 Language Testing

This part will be divided into several parts, which are the History of Language Testing, Definition of Language Testing, and the Validity.

2.1.1.1 History of Language Testing

Language testing as a methodology for analyzing and investigating language proficiency has been around a long time and comes from the traditions of human's need to have teaching and learning activities. Being as one part of language teaching, it provides a goal for language teaching and it monitors progress, to be used by teachers and students in achieving its goals in the future. Language testing also provides a methodology for experiment and investigating, both languages teaching and language learning.

As Spolsky (2001) say, in the first 2000 year in human history human

abilities has begun to be assessed formally, and this makes tests and examinations are in a strong position at that time. A century ago, Examinations judged as not having a clear and obvious certainty, but the testing practices (teacher, government, etc.) at that time still managed to keep it growing. Nowadays, the appreciation of the fundamental idea of the test as "language proficiency", and public acceptance of the reality of the impossibility of interpretation done on one measurement alone, led to the realization that the measurement of language is actually more complicated than previously imagined, that could not be shown by a simple score on paper.

According to Davis (1990), the thing that makes language testing became important in Applied Linguistics is, unlike other subjects that offer education and education to students, a language has no clear content. Therefore, assessing the complexity of language produced what should be tested, or aspects of the exam itself as the problem of testing the validity of the instrument, how the test is given, what should be tested, and so forth.

According to Shohamy (2001), tests are often used as a measurement tool designed to obtain specific data either directly or indirectly. In fact, as a tool used in the education system, the tool is used in large scale and has an extremely powerful force.

Based on Confucian doctrine during the Han Dynasty as the initial occur of formal testing, Splosky (2001) said that we have a long history to explain our understanding of the whole development process of assessment until it becomes the current form like now. He further states five major purposes for test

using as mentioned below:

- Using tests as a competitive selection device
- Using tests in order to provide information on the quality of the "product" to those who are paying for an education system
- Using tests to process and certify that an individual has achieved a specific level of technical or professional skill
- Using tests for prediction or prognosis of the probable results of training
- Using tests as an integral part of all good teaching

Arguably the most common usage of language test is to highlight strength and weakness in the learned abilities of the students. Based on Farhady et al. (1996) tests are applied to make decisions about people's lives in general terms, so fair decisions will be impossible to achieve if the test do not provide clear and complete information. On the other hand, specific behavior samples can be obtained by test, which differentiate it from the other types of measurement (Mousavi, 1999). So we can say that any procedures or activity to assess and measure particular ability can be called as a test.

Language testing as part of applied linguistics indeed has changed and expanded in various ways in recent years. Bachman (1999) presents a brief review in language testing, that he examined in 1977 that the practice of language testing is given in terms of just where the theoretical view of language proficiency skills (listening, speaking, reading, and writing) and its components (grammar, vocabulary, pronunciation). This approach makes the exam at this time isolated in

fragments of language elements only, not in the assessment form of overall language ability. Language testing research also dominated mainly by the hypothesis that language proficiency consists of a single trait only, and using a quantitative methodology as a means of measuring statistics.

In particular, works of applied linguistics such as Widdowson (1983), Savignon (1983), Canale and Swain (1980), and many other prove that language testing subjects is very influential and beneficial for many people and many aspects. Their view that the use of language in part triggers the initial conversation, as the search for meaning, or as a dynamic open many people's minds that language ability can not be isolated in one aspect, but must relate other aspects such as aspects of language in terms of sociolinguistic as well as consideration of the context in which in question took place (Bachman, 1999)

In 1980, language testing has wider coverage than ever before. Most important perhaps, under the influence of SLA research, led language testers to investigate not only factors such as talent, influence, dependence of the field, and conversations in the language domain test performance, but also the strategies involved in test-taking process itself (Bachman, 1983). Whereas in 1980, awareness of the importance of language testing as part of applied Linguistics, as Bachman (1999) defend, in year 1990 this trend continues and accompanied by research methodologies. In this year also the relationship between language testing and educational measurement are connected, particularly in areas where technical process of methodological, ethical aspects and the consequences the test. It then continues to grow up so in last decades language testing methodological

approaches become more diverse, specific, and include more aspects that are believed to influence the overall outcome of these studies.

In 1990 the growing use of technology to the level of development and delivery of test forms. Progress in this cases it possible to test the ability of students to be able to adjust in line with the increased access to computers and the level of consciousness in the technology. This makes the test run large-scale computer technology is growing rapidly, and making the experience of participating tests and administrating as well as monitor the whole process much more easier than before.

Changes that occurred throughout the history of language testing, making Spolsky (1978) and Hinofotis (1981) divided it into several categories of time. Hinofotis (1981) calls this change as a trend, and then divide it into three trends, namely the pre-scientific, psychometric-structuralist period, and integrative sociolinguistic period. Brown (1996) also mentions the existence of such classification, but chose to mention this change as a movement rather than a period, as changes occur somehow overlap and are still widespread in parts of the world until now.

2.1.1.2 Definition of Language Testing

Language testing is a full measurement process that is performed, so that people who used can be able to measure how students' ability precisely and thoroughly involves in the process of language teaching process. The existence of social dimension of assessment might well be more striking in language testing

than in assessments measuring general cognitive abilities because language is a social medium and any measurement of it outside of a social context tends to be at odds with the increasing acceptance of social models of language within Applied Linguistics more genetically. Of course, language tests for a long time ignored the social use dimension of language and followed traditional psychometric methods of measuring isolated pieces of grammar and vocabulary knowledge. However, with the rise of the communicative competence paradigm in second language teaching, the measurement of the ability to use language in social contexts has become increasingly important (Blackwell, 2006).

Language test can be a valuable tool for providing information that is relevant to several concerns in language testing. As Bachman stated (1990), language testing can provide evidence of the result of learning and instruction, and hence feedbacks on the effectiveness of the teaching program itself. They can also provide information that is relevant to making decisions about individuals, such as determining what specific kinds of learning materials and activities should be provided to students, based on a diagnosis of their strengths and weaknesses, deciding whether individual students or an entire class are ready to move on to another unit of instruction, and assigning grades on the basis of the students' achievement

Finally, language testing 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 used needs of the students following the program of instruction. For

these reasons, all language teaching programs involve some testing, and hence, language teachers need to be able either to make informed judgments in selecting appropriate language test or to plan, construct, and develop appropriate test of their own (Bachman, 1990).

2.1.1.3 Validity

According to Brown (2003), a good language assessment has principles to do before it is considered as a good assessment technique. It is also necessary to make a test based on the assessment principles in order to make it reliable and better.

Arguably the most complex and complicated principle for an effective test is Validity. Gronlund (1998) describe the validity description as "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment". A test that considered valid is a test that measures certain ability, which the test wants to seek from the students or audiences (Brown, 2003). So if a test wants to measure a writing ability from a student, the test should test the writing ability only and not the others. The idea of validity is to measure the exact item the test wants to seek, and give a clear explanation to the students and how the question and task in the test will measure that ability.

There are no exact and certain methods used to measure the validity of a test. But some certain method that has been used over the years have some sort of evidence that will keep the validity of the test, such as to test the students with

the same unit they have been learned in the class. There is also a consideration about the test consequences that focus on the aspects beyond the test itself such as the effect of the test taker or perceptions towards the test's validity. There are five items of validity that will be explained:

2.1.1.3.1 Content Validity

Content validity is a term where a test requires the students to perform an ability that will be measured on the test directly. Generally, we can identify if a test contains a content validity, if we can clearly define what is being tested within the test, and if the test's final achievement is the same with the task that is requested by the test itself. A test is considered not having a content validity if the test asks the student to do something else different from what the test want to seek from students.

Brown (2003, pp. 22-23) give an illustration for this content validity by this example: a tennis competency that requires someone to perform a 100-yard dash, measuring certain person's ability in speak a second language in conversational setting by making the learner to answer written multiple-choice questions full of grammatical judgment, a test that want to see a person's fluency in speech by making them to mention words as many as they can; those examples are considered lack in having a content validity.

If the students had given a material about certain topics, and finished practicing the materials in many form like in written, listening, and speaking mods; and then a question is given to them in written form containing similar

setting with the materials to test their understanding towards what have they learn, then this type of test is contain a content validity.

The content validity can also be understood by how the test is being delivered to the student. There are two types of the techniques that are direct testing and indirect testing. The direct testing requires students to perform the task in the questions, while the indirect test doesn't test the task itself but the other task that will still related to the main task. As long as the other task will still give evidence on how well students perform in the main task requested, than it is still regarded as having the content validity.

The idea of content validity and all the examples explained above is, it is not the only type of validity, but teacher and anyone who may making a test should look up about validity. Content validity doesn't require as high observation within itself to look that a test contain this validity or not compared to other types of validity, and it is also considered as the first evidence that would be looked up from a test, and teacher or anyone who would make a test should keep this content validity as their priority.

2.1.1.3.2 Criterion-Related Validity

Criterion-Related Validity shows that a test has this kind of validity if it considered being accurate showing a result with another type of test in same domain by comparing both of them. An accuracy of its result would be looked together and checked if it resembles the other test's results too. And thus, this can be categorized that a test has a criterion-related validity.

For example, student who participates in intermediate music theory exam in his college, and he scored above average grade. Thus this score will be checked with the practice test conducted with the same academy. The score from the written test that will be checked and compared with his practice test score, is the practice score is as good as his written test. If the result from both test are similar, then the written test has the criterion-related validity within it. This comparison also can be used to indicate how a test is correlated to another similar test, and how the test affects each other on each domain.

Criterion-Related Validity can be divided also into two types, which is Concurrent Validity and Predictive Validity. Concurrent Validity is type of criterion-related validity that obtained criterion measurement at the same time as the test scores. Concurrent validity indicates to what extent a test would accurately estimate an individual's current condition and state with its result. The example of this would be a test that would measure a proficiency of English fluency. The test result should show the same (or close) as the speaking test, as another valid method to measure someone's proficiency beside the written test.

While Predictive Validity occurs when the measurement result are obtained some time later after the test has been conducted. The best illustration of predictive validity is an aptitude test or career test that will help to see what subject a person excel the most among the other skills he have, and what job he is suitable in the future. The time that the result of both validities obtained is the thing that distinguishes concurrent validity and predictive validity.

2.1.1.3.3 Construct Validity

Construct Validity defined as any theory or hypothesis that explains an phenomena with the understanding of general perceptions. In language assessment, construct validity could be measured whether directly or indirectly. Because construct validity served as the theme, we could say the general idea of the whole aspects of teaching. Davidson et al (1985) described that a test are "operational definitions of constructs in that they operationalize the entity that is being measured".

The construct validity in language assessment appears as the theoretical construct. Thus means every test that has been made should be refer to the theoretical side, that every point that would be measured still lies in the corridor of how should the ability being tested by general justification.

Illustrated by speaking test conducted by a teacher, and the scoring point of the test will be measured from pronunciation, fluency, grammatical accuracy, vocabulary use. The decision to use this point lies in a theoretical construct that accept those ability point described before will be the main skill used in the real life usage; thus it would be considered valid that using the factors as the measurement item in a speaking test.

Generally construct validity will perform well in test that handle small to medium scale of student, as with the large scale the time needed will be need a lot of time; and the practicality side of the test is omitted that way. Also, large scale test need to consider the financial side of the test's execution, that should be prioritize that the test will perform well with that limitation.

2.1.1.3.4 Consequential Validity

The concept of consequential validity had its initial movement when Messick (1989) introduced it as one of the validity concept that must be included in a test. Later this concept also strengthened by Shepard (1993, 1997), stating that both positive and negative effect emerged from a test should be analyzed more as a part of validation system of an assessment, and pushing the limitation of validity concept to sought after the external factors too.

Consequential validity covers items that appear as consequences of a test, including the test's accuracy to measure the in related category, the effects towards student's preparation, and the subsequent effect to student's learning progress, intended and unintended social consequences from interpretation to the test itself (Brown, 2003, p. 26).

The intended goal from assessment are set to target mainly the students and teacher's in their implementation of their learning progress, such as curriculum and instructional content and strategies, classroom format and types of material used due to the test type, and also their motivation and effort towards the learning and teaching process. Also how the test performs should change public and people perception towards the test (Lance and Stone, 2002). However, some unintended consequences may occur as a result such as eliminate unnecessary material that may come out on later test, or using a material that closely related to the test without making any modifications.

In 2012 National Exam, The minimum limit for each subject score

was 4.00, and the minimal average of the whole subject is at least reach 5,50. Due to the latest release in *KTSP* curriculum give the schools the right to design the curriculum and the learning models to be applied, and then certainly there will be two different standards of value, which are the standards set by Indonesian government and the standards, came from the school. And in this case, the respective high school determine its minimum score limit, which is 77 (7.7).

From here, we can know that the standards have been set by the schools is much higher than the standard set by government, as has been implemented in the national exam. We can conclude that students of SMA Negeri 1 Sidoarjo has a very small chance to obtain score below the national standard, because they have been accustomed to the higher standard set by their school.

With the explanation mentioned above, then the separations of students who have score below average with the students who have a score above average based on the national standard become impossible. Therefore, the limit that will be used to put students into different groups is the one from the school, which are 7.7. Respondents' score in the English exam will be collected and then the median of those score will be set. The later median will be used as the border of grouping the respondents. Respondents that put in the group of student with below-average score are students that having score below the median, as work the same to with the students with having scores above the median; will be put in the group of student will above -average score.

2.1.1.3.5 Face Validity

A test can be said to have a face validity if it looks like would measure what supposed to be measure. It is the value of the test itself that looked up by people, and comment if it is acceptable as a proper test to measure certain ability. If a test considered as a valid test, then test-maker may use the data and determine which kind of test that is a valid in the future as a consideration. The face validity also important, as it will directly affect the student's behavior towards the test.

Mousavi (2002, p. 244) defines face validity as "a degree which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers".

This is why validity is considered important, especially to let the students know about the test, in order to avoid the student's lack of understanding about what the test would measure, or which area the test wants to measure. The understanding comes from the students about the face validity of a test is believed to made a test appear good in front of them; thus making the test halfway point away as a full valid test.

Brown (2003, p. 27) gives some example of point that could be used as judgment of a test to have face validity, are:

- Well-constructed format with familiar tasks that students have encountered before
- Can be finished well with the time restriction given

- Material that is not complex and easy to understand
- Directions in the test is no confusing
- Tasks that is related to their lesson work
- Difficulty level that is reasonable to them; regarding what have they learned

Face validity is purely an external factor that cannot be controlled by the test-maker itself. It is cannot be tested empirically and with theoretical field, and just how Stevenson (1985) view this validity as a "factor that is dependent on the notion of the perceiver". Because not everybody thought how the test should be delivered in particular way, the opinions from the students who see the test are very important. If some students feel that the test is not appropriate to test a specific matter that they have learn, then the test is already fail one of its purpose, to be a good measurement test on specific ability.

Generally face validity is achieved by inserting a good content of material into the test, and recognized as a good way to test the material too. Test that has content material the students had studied or expect to appear with the item they have learned, and then most likely it has face validity. A good content validity will delivers a proper and exact material with the students have learned before; therefore this process will also eliminate their anxiety and hesitation, claiming whether a test has a face validity within it or not.

2.1.2 Indonesian National Exam

Indonesian National exam is a standard evaluation system of primary

and secondary education nationally conducted by the Educational Assessment Center of Ministry of National Education in Indonesia (Undang-Undang Republik Indonesia No. 20 Tahun 2003), states that in order to control quality nationally, then education evaluation is done as a form of accountability of education providers in Indonesia.

The national exam serves for purpose to achieve competence standard of the students for particular subjects in groups of science and technology, based on educational system. The national exam held on the fourth week of April, which is the main exam that held together nationally in all schools in Indonesia, and the next week of it the same exam also, but just for students in special case who cannot join the national exam a week before (e.g. sick, special permission, etc.) (Permendiknas No. 34 Tahun 2007)

The purpose of national exam is used as one of the consideration for several purposes and uses, based on the *Permendiknas No. 34 Tahun 2007*:

- Mapping Units and the quality of educational program
- Selection into the next level of education for the students
- Determining the graduation of students from the program and the education unit
- Coaching and providing assistance to the education unit in an effort to improve education quality

In the real application, the person who in charge of the national exam is the Indonesian Minister of Education, who responsible of all aspects included in the process of holding the national exam, and helped by the other national

agencies below the Minister himself. The minister responsibility prior to several points, such as stated schools for the students in Indonesian schools that located outside Indonesia's country area (placed in another country), stated the total cost that is needed to do the whole process of national exam inside and outside the country, preparing the *SKHUN (Surat Keterangan Hasil Ujian Nasional)*, and finally to monitors, evaluate, and establish the subsequent action of post-national exam.

2.1.3 Perception towards National Exam

Theoretically, perception is described as interpretation of information to create a mental representation of the environment by converting the signal into nerve signals that can be understood by the body (Schacter, 2011). There are two methods used to retrieve the information into the perception known to our bodies; the first is the top-down method and the second is the bottom-up method (Goldstein, 2009, p. 10). The top-down method uses the person's experience and knowledge to creating the image into the perception of particular object while the bottom-up method use many information about particular object, construct and combine it all into one image of perception. Both of the method may be used simultaneously in real application, as more information is generally better in to create image that become a perception later (Bernstein, 2011). Perceiving is a complex process that blends the imagination from the person, the thing that perceived, and the surrounding factor at the time.

As Reeder et al. (2012) stated, perception towards a test built based by

the participants of the test taker (students) and the test form that combined into a face validity of the test. The factors are students' previous experience about the test and their belief about their ability. Factor that emitted from the test is about the clearness of the test that wants to assess particular item of a material.

In the study, they gather two types of participants, a student with academicals background and community through flyer. Then these two groups are given two types of test for simulation of job seeking test, that is numerical test and specific subject test; in this case machinery knowledge test. From their study, the ability that come out and influenced by perception are most likely come out from their own believe about their ability, such as whether they believe they can finish the entire test with a good score, or their motivation about passing a test. The factor that may influence face validity from test aspect is that the type of question. It is noted that a test that formed with many focused question about material gained higher face validity than the test with general type of question, such as numerical test in the study as numerical test have many application in different setting of a job.

Thus perception that create from the test would be focused on how the test is clearly understandable to the test taker such as what it supposed to test, how effective the type of test of would assess particular ability, and the test's similarity with the material being tested.

In terms of this research, the elements of the perception that is related to what extent the test is understandable to the test taker, what it supposed to test and its similarity to the material being taught in the class and during the

preparation is used in the questionnaire. Their ‘perception’ is documented in this study using the collection of data and information from any possible sources originated from their experience, for example their comment about the appearance of the test, previous students’ comment about the national exam, previous possible resemblances that may occur in the process of perceiving, past memories about similar test and experience, and other more. External factors also added in this study. Report about national exam in national newspaper and online Medias, case of cheating and flaws during national exam execution, social comment and behavior towards the test would also likely be a consideration in creating their perception.

2.2 Related Studies

As mentioned briefly before, on example of study that sought the relationship between perception and achievement entitled "Reactions to Cognitive Ability Tests: The Relationships between Race, Test Performance, Face Validity Perceptions, and Test-Taking Motivation" (Chan et al, 1997). This research studied how the relationship between related variables is suspected to influence the final results of test takers.

The results showed that the perceptions and student motivation is positively related to their performance in such tests, even after the effects of race and performance aspects of the first test has been controlled. Effects resulting from racial aspect in this case is more likely the differences in cultural background relating to the perception of their tests, and it can be said almost have

no direct relationship arising from racial aspects of the students. And also perceptions of face validity here affect subsequent performance tests, but only affect a portion of the overall results, and this can be seen from their motivation in following tests. This shows the influence of external factors other than the test itself also influence the actual performance of students in their overall performance.

Another study titled ‘Predictive Validity in the IELTS Test: A Study of the Relationship between IELTS Scores and Students' Subsequent Academic Performance’ (Kerstjens & Nery, 2000), sought about the impact of the IELTS test is examined to predict academic performance of students in Australia. In this study, some students were given a questionnaire to obtain data about the perceptions of students, which is used to see the validity of the IELTS (in this case the predictive validity). In the questionnaire, respondents were given a question that will be used to determine whether their English language skills good enough to take the first semester of their studies, the difficulty level of IELTS according to them, and language support they received during the first period of their study that are given simultaneously. In the end, the research shows that the perception of students is also one of the important things in order to test the validity of a test.