CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter provides some theories of literature related to discussion of the study. Also, it presents review of previous study to show the differences between this research and other previous researches which is done by other researchers.

A. Review of Previous Study

Before doing this study, the researcher has read some previous studies focused on the same topic as it does. The first previous study is the thesis from Ita Faradillah entitled *An Analysis of Essay Test on English Final Test for Grade Eleven Students of SMAN 1 Lamongan*[1]. This study investigate the content validity, index difficulty and item discrimination of essay test tested in two different classes, XIA5 and XIA6. The result showed that the essay test has good content validity which is proven by 80% of concurrence. It means that most of test items represent all material taught in grade eleven although there are some items which are out of the box. Then, the index difficulty and item discrimination of essay test produced different result in those classes. It was acceptable only in XIA5 but it was rejected in XIA6 as the result is around 0,1-1,0 for index difficulty and 00-0,19 for index discrimination which means that the items were

[1] Ita Faradillah, thesis *An Analysis of Essay Test on English Final Test for Grade Eleven Students of SMAN 1 Lamongan*. State Islamic University of Sunan Ampel Tarbiyah Faculty of English Education Department. (Surabaya: State Islamic University of Sunan Ampel, 2012)

too easy and too difficult. Therefore, the essay test should be revised for XIA6 because it could not discriminate students' achievement properly.

Unfortunately, this study did not measure the level of raters' consistency in scoring essay test. Moreover, a good test must complete the requirement of validity and reliability. This study only investigates the validity not reliability. So that in this study the researcher wants to examine the reliability, look at from raters' consistency, to complete this previous study.

Second, *Classroom Writing Teacher's Intra- and Inter-rater Reliability: Does It Matter* by Viphavee Vongpumivitch[2] from National Tsing Hua University investigated the reliability of each scale in an analytic scale of scoring essay test and the teachers' consistency in scoring essay test using the scale. The result showed that the teachers have very low intra- and inter-rater reliability; each scale of the analytical rating scale has low correlation, especially in content and organization; and each teacher has different understanding of the scale's criteria based on their experience, personalities, and personal agendas. This study had measured the inter- and intra-reliability of each rater thoroughly in using rubrics (scale's criteria) for assessing writing.

[2] Viphavee Vongpumivitch. *Classroom Writing Teacher's Intra- and Inter-rater Reliability: Does It Matter*. Journal of International Conference on English Instruction and Assessment National Tsing Hua University, 2006.

The next is *Rater Discrepancy in the Spanish University Entrance Examination* by Marian Amengual Pirazzo[3] from University of Balearic Island. This study told that there are no significant differences between the holistic pre- and post-scores but there are important differences in the behavior of raters in consistency of scoring. In short, the intra-rater reliability is quite high despite some exceptions such as their condition in scoring, etc.

The last two studies have examined the intra-rater reliability of scorer in scoring essay test. It has measured the raters who use both holistic and authentic assessment. As this is the first research of intra-rater reliability in Indonesia, especially in English Education Department of UIN Sunan Ampel Surabaya, so that it will measure the intra-rater reliability in general. The researcher makes it special as the subject of this research is the English teachers of Al-Amin Islamic Boarding Senior High School Mojokerto where the English learning focuses on the students' language skills, especially in writing.

Another previous study is *Reliability and Validity of Rubrics for Assessment through Writing* by Ali Reza Rezaei from California State University and Michael Lovorn from The University of Alabama, USA.[4] This study intended to investigate the reliability and validity of rubrics in the assessment of students' writing prompt. The results showed that rubrics may not improve the reliability

[3]Marian Amengual Pirazzo. *Rater Discrepancy in the Spanish University Entrance Examination*. Journal of English Studies University of Balearic Island Vol.4 page 23-26, 2003-2004.
[4] Ali Reza Rezaei – Michael Lovorn. "Assessing Writing". *Reliability and Validity of Rubrics for Assessment Through Writing* Vol. 15, 2010.

and validity of assessment if raters were not well trained on how to use and implement the rubrics. This result rejected the writers' hypotheses that the rubric will improve the raters' reliability and validity as it was more analytic. It was proven as the raters were more influenced by mechanical characteristic than the content even they used a rubric.

Even this study uses a rubric for scoring essay test but the focus was not the rubric's impact in the raters' assessment. This research only focuses on the raters' consistency of intra-rater reliability in scoring essay test based on all categories in rubric, such as Content, Organization, Grammar, Vocabulary and Mechanic.

B. Theoretical Background

1. Understanding of Consistency

Based on some dictionaries, consistency is the ability to remain the same in behavior, attitude or qualities. It means that consistency identifies how stable someone in doing something. In this case, consistency analysis of English language teachers in scoring essay test means measuring the quality of human raters; English language teachers, in rating their students' essay test.

The consistency level of reliable test can be estimated by calculating a reliability coefficient ($r_{xx'}$). A reliability coefficient is like a correlation coefficient that the maximum is +1.00 for a perfect reliable test. It can be

interpreted as the percent of systematic, or consistent, or reliable variance in the scores on a test.[5]

According to James Dean Brown, reliability coefficient is different from a correlation coefficient in that it can only go as low as 0.00 because a test cannot logically have less than zero reliability.[6] Therefore if there is negative for the reliability of the test, he suggests checking for errors whether the researchers make mistakes in their calculation. In addition if the calculation is right, they should round the negative result to 0.00 and admit the results on the test have zero reliability which means totally unreliable or inconsistent.

2. Understanding of Intra-rater Reliability

One of ways to measure the quality of good test is reliability. To be reliable, a test must be consistent in its measurement. In other words, a test score has to be free of measurement error.[7] It is like when the teacher gives the same test to the same students on two different occasions, the test should produce the same result too. If it is different, it has many possibility factors; such as it comes from the students, the examiner or rater, the condition when

[5] James Dean Brown, *Testing in Language Program*….. 175.
[6] Ibid., 175.
[7] Daniel Mujis. *Doing Quantitative Research in Education with SPSS* (London: Sage Publication, 2004), 71.

the test happen (test administration) or the test itself (the length of the questions, the paper used, etc).

The student-related reliability is some factors that may affect reliability that come from the audience of the test, mean the students. It is caused by temporary illness, "bad day", anxiety, and other internal physical and psychological factors of students.[8] Whereas the test administration reliability is the reliability factors that include the condition during testing process, such as the noise street so the student who sit beside the window cannot hear the tape recorder clearly in listening test, photocopying problems, the amount of light in different part of the room, temperature problems or the condition of desks and chairs. Then another factor that can affect the reliability comes from the test itself. It can be caused from the length of the test is not balance with the time longer.

The last factors that may influence reliability come from examiner or rater of the test. There are two kinds of rater reliability. First is inter-rater reliability, when there are two or more scorers examine the same test. If they produce different score, it means the test has low-reliability or unreliability. Second is intra-rater reliability, it is not about two or more scorers but only one scorer examine the test individually. This type has bigger possibility of subjectivity because of unclear scoring criteria, fatigue, bias, simple carelessness or sympathy to students.

[8] H. Douglas Brown, *Language Assessment*..... 21.

Rater reliability, especially the intra-, is one of repeated measurement reliability form that conceptualized in quantitative research. It has to do with the ability to measure the same thing in different time, called test-retest method. By using this method, this study wants to investigate how strong the relationship is between the scores at the two time points, in this case is consistency. The intra-rater reliability coefficient can be resulted from the average or the added up of two sets of scores in the decision making process. Cronbach Alpha is the most appropriate form to calculate this coefficient. It is the easiest formula in split-half reliability method. The result will be formed in decimal and it will show the reliability level of each teachers. In addition, it will help to give the final result whether the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto consistent or not in scoring essay test.

One important thing that follows when measuring intra-rater reliability is how much time needed to let go by before post-scoring. This is very difficult to answer as every research about this topic has different time interval. For example, the journal entitled *Writing Teacher's Intra- and Inter-rater Reliability: Does It Matter* by Viphavee Vongpumivitch[9] from National Tsing Hua University had one week interval between pre- and post-training stage whereas another study entitled *Rater Discrepancy in the Spanish*

[9] Viphavee Vongpumivitch. *Classroom Writing Teacher's Intra- and Inter-rater Reliability: Does It Matter*. Journal of International Conference on English Instruction and Assessment National Tsing Hua University, 2006.

University Entrance Examination by Marian Amengual Pirazzo[10] from University of Balearic Island gave the distance for pre- and post-scoring in three months interval. If the time interval was too short, the raters may remember how they scored last time and simply give the same score because of this. In contrary, the raters' opinion may be genuinely changed. It is called carryover effect and can lead overestimating the reliability of the test.[11] One to two weeks is often recommended as an optimal time, though the risk of some carryover effect remains.[12] To reduce and avoid the risk of carryover, this study used two months interval as it was not too short like one week and too long like three months.

Unreliability or inconsistency is clearly a problem. Inconsistency rater will lead to unreliable test that can influence the score produced. The score may be impacted by many factors which indicate the unreal grade, means it does not represent students' real condition. Therefore students will not get the appropriate feedback and mark based on their true ability.

This research only focuses on intra-rater reliability of the test rater. It is the most suitable reliability that should be researched as the condition of English teachers in Indonesia who teach the class individually. Therefore, the researcher wants to know the consistency of English teachers in scoring essay

[10]Marian Amengual Pirazzo. *Rater Discrepancy in the Spanish University Entrance Examination*. Journal of English Studies University of Balearic Island Vol.4 page 23-26, 2003-2004.
[11] Daniel Mujis. *Doing Quantitative Research*…. 72.
[12] Ibid., 72.

test as subjective test in the grade eleven of Al-Amin Islamic Boarding Senior High School Mojokerto.

3. Understanding of Essay Test

According to Coffman, essay test is one or more questions administered to a group of students under standard conditions for the primary purpose of collecting evaluation.[13] Essay items are useful when teachers are interested in learning how students arrive at an answer as they do not ask students to choose one of responses like objective test but to share their ideas by their own word. In this test type, students decide how to approach the problem, how to set it up, what factual information or opinion to use, and how to specifically express their answer.

Based on Stalnaker's definition, an essay test should meet the following criteria:[14]

- Requires examinee to compose rather than select their response.
- Elicits students' responses that must consist of more than one sentence.
- Allows different or original responses or pattern of responses.
- Requires subjective judgment by a competent specialist to judge the accuracy and quality of responses.

[13] W. E. Coffman, *Essay Examination*….. 271.
[14] Christian M. Reiner – Timothy W. Betwell, *Preparing Effective Essay Questions: A Self-Directed Workbook for Educators*, (New Forum Press: 2002), 6.

Essay test is the most appropriate part to measure students' cognitive skill because it explores students' critical thinking and conscious mental process. Age is the influence for human cognition, it develops rapidly throughout the firs sixteen years of life and less rapidly thereafter.[15] Therefore this research concerns on grade of eleven, as the participant of the essay test that taken the score, which the average of students in the class are sixteen to seventeen years old. In addition, the essay material is focused on the grade eleven so that the researcher is sure that all students have gotten the material well and they will give their best in doing this test.

Nowadays, most of English teachers have increasingly turned away to this essay test. They have some motives why they choose it than multiple-choice (MC) test. Moreover, the assessment of MC is easier than essay as it is one of kind of objective test. The reasons are:[16]

- Assess students' higher-order or critical thinking skill, means that this test can test complex learning outcomes that cannot effectively assessed by other assessment procedures.
- Evaluate students thinking and reasoning, means that this test can examine thought processes from how the students select, organize, and evaluate facts, ideas, etc.

[15] H. Douglas Brown, *Principles of Language Learning and Teaching Fourth Edition* (San Francisco State University: Longman Inc., 2000), 61.
[16] Christian M. Reiner – Timothy W. Betwell, *Preparing Effective…….* 10.

- Provide authentic experience, means that this test can assess the students' ability to construct solution and decision.
- Require students use own writing skill;[17] the students can select their own words, sentences and paragraphs or organize correct grammar and spelling.

Besides the strength, there are many weaknesses which are contained in the essay test such as the lack of validity and reliability, the unpredictable result, difficult to assess and the longer time needed to examine. The main problem of this test is seemed from the examiner reliability. The testers get many difficulties in deciding the score of each essay test. Even there is a rubric which contains scale criteria but there is still subjectivity during scoring process. In addition, English teachers in Indonesia are never be aware of their objectivity in scoring essay test as subjective test. They score the test individually without caring whether their score will be stable or not when they try to score in different occasion. Therefore, there is a big possibility that their assessment of each essay test will change based on the raters' internal or external situation when they score the essay (low intra-rater reliability).

There are two types of essay test, extended and restricted response question.[18] It is distinguished from the choice of the content and the form. Extended allows students to decide the content and the format freely. While

[17] William E. Chasin. Idea Paper No.17: *Improving*….1.
[18] Ibid., 1.

the restricted limits students in choosing both of them. Most writers agree that this type is the most appropriate form when the teachers wish to test content. This study uses restricted response question in form of exposition essay as this essay test is one of daily examination which is held to know students' achievement in a specific material based on curriculum. Therefore, the teachers can examine each students writing ability clearly.

4. Understanding of Scoring Essay Test

As essay test is a subjective test that has subjective nature and complex judgment, this assessment has gotten big attention, especially in human raters. Even if there are many applications offer automated scoring of essay test but human still have a big part in this assessment as they can understand both the content and the quality of writing. Some of the strengths of scoring by human rater are that they can (a) cognitively process the information given in a text, (b) connect it with their prior knowledge, (c) be based on their understanding of the content, make a judgment on the quality of the text, and (d) be able to recognize and appreciate students' creativity and style.[19]

Beside all the strengths, human scoring has limitation. Some of their weaknesses are needed good human rater quality and instructed in how to use

[19] Mo Zhang, *Contrasting Automated and Human Scoring of Essays*. (Article of R&D Connection No.21 March 2013) www.ets.org , 1.

scoring rubric so that they must be controlled continuously.[20] In addition, they can make mistakes based on the cognitive limitation which is difficult to quantify and cause systematic bias of the score.[21] This bias can make the validity and reliability of the essay test automatically low. Therefore, it is very important to check them immediately. As there was study about examining the content validity, the index difficulty and item discrimination about essay test so that the researcher will examine another part of essay test. This study will measure the reliability of the essay test observed from the individual grader of the test, intra-rater reliability.

There are two tools that can be chosen by testers in scoring essay test, holistic (global) and analytic (point-score) assessment. Its evaluation and description become the main difference between those tools. The analytic allows for separate evaluation of factors to be evaluated (e.g., persuasive argument and grammar in writing) and the description of what is expected at each score level is provided. In short, the analytical assessment is characterized by a specific scale's criteria which decide how much of each maximum subtotal judge the students' answer to have earned.[22]

While the holistic is used when it is not possible to separate an evaluation into independent factors, or if there is an overlap between the

[20] Ibid., 2.
[21] I. I. Bejar, (2011). A validity-based approach to quality control and assurance of automated scoring. Assessment in Education: Principles, Policy & Practice, 18(3), 319
[22] William E. Chasin, Idea Paper No.17: *Improving Essay Test* ….. 3.

criteria set for evaluation of different factors. In addition, it supports broader judgments concerning the quality of the process or the product. Simply, this tool is indicated by a whole evaluation which makes an overall judgment about how successfully the students have covered everything that was expected in the answer and assigns the paper to a grade.

English teachers in Al-Amin Islamic Boarding Senior High School had used analytic assessment in scoring an essay. They have a rubric but it was too general. There were no specific criteria for each grade. Therefore, the teachers were helped by a rubric to help them to be more detail. Since this study was the first research about rater reliability of essay test in Indonesia, especially intra-rater reliability, it ignored the assessment tool used by the teachers.