

CHAPTER I

INTRODUCTION

This chapter presents the background of the study to explain some important reasons why the researcher conducts a research about this topic. Then the problem is formulated in research question with the objective of the study. The usefulness of the research is mentioned in significance of the study to show how important the research is. In addition, scope and limitation explain the constraint of the study that some parts have to be ignored here. Also it is completed with definition of key term which gives simple descriptions of some important key words of the research.

A. Background of the Study

Today, ESL teachers or tutors (English language program teachers) more decide to use performance-based assessment in evaluating students' language ability, such as oral production, writing production, open-ended responses, integrated performance (across skills area), group performance and other interactive tasks.¹ It can be called as direct assessment since it provides more direct evidence of meaningful application of knowledge and skill.

¹ H. Douglas Brown, *Language Assessment: Principles and Classroom Practices* (San Francisco State University: Longman Inc., 2004), 10-11.

There are many kinds of writing test, such as paragraph construction test, short-answer and sentence completion test, picture-cued test, essay test, etc. This study will be focused on the essay test as the most appropriate test to measure students' critical thinking and conscious mental process. Essay test is one of writing production evaluations that requires the student to structure a rather long written response up to several paragraphs. Their scoring requires expert judgment rather than the application of a clerical key.² This test is commonly chosen to evaluate students' writing ability since it has strengths, such as it can test students' complex learning writing outcome, posing a more realistic task and also requiring students to use their own writing skill.

While the fact that constructed-response essay items ask students to produce samples of normative language which make such items to have more valid measurement of communicative writing ability, the process of scoring essay items is quite complex. In this part, human raters experience a new defiance to defend the reliability and construct validity of test scores, as each rater has different perceptions of performances and tendencies of leniency and severity. It can cause high possibility of subjectivity as raters have their own personal criteria of scoring.

² William E. Chasin, Idea Paper No 17: *Improving Essay Test* (Kansas State University: Centre for Faculty and Development, 1987), (5429-5191-887-10M) 1.

The subjectivity in scoring a test can cause a doubt on the stable of estimation. According to the explanation above, not only subjectivity but also human error and bias during scoring process can deliver inconsistent grade of each student. This inconsistency can decrease the rater reliability that can reduce the assessment accuracy. In evaluating writing ability, rater reliability is difficult to achieve since writing proficiency involves numerous traits that are hard to define.³ It is the same as James Dean Brown statement that the very difficulty of measuring mental traits explains why consistency is of particular concern to language testers.⁴ It means that the consistency of scoring is very important to be aware by all English teachers or testers since it determines the objectivity of the “real” score that each students should be gotten.

The rater consistency in scoring essay test can be measured in two ways, inter-rater and intra-rater reliability. Inter-rater reliability is estimated by looking at the scores produced by two raters and calculating a correlation coefficient between the two sets of score.⁵ Many foreign researchers had more chosen this type to evaluate rater’s stability as the objectivity is more visible from comparing

³ H. Douglas Brown, *Language Assessment* 21.

⁴ James Dean Brown, *Testing in Language Program: A Comprehensive Guide to English Language Assessment* (Singapore: The Mac Graw-Hill Companies, Inc., 2005), 169.

⁵ *Ibid.*, 185.

scores produced by two or more raters. In fact, English teachers in Indonesia are not accustomed to teaching and assessing their students' test in pair.

Intra-rater reliability is paid big attention here. It is a common occurrence for a class teacher, like teachers in Indonesia, because unclear scoring criteria, fatigue, bias toward particular "good" and "bad" students, or simple carelessness.⁶ This type allows us to measure the individual rater by comparing two scores produced by the same rater from the same essay in different time (pre- and post-score). Unfortunately, Indonesian English teachers do not have big attention and awareness in evaluating their consistency in scoring their students' test, especially essay test as one of subjective test that the raters may have subjective nature during scoring process. It is proven by the unavailable studies which try to examine the rater consistency using this method.

The researcher has made a survey of theses in term of language testing which are especially done by students of English Teacher Education Department UIN Sunan Ampel Surabaya in their thesis. They mostly have concentration in some assessment types, such paper project assessment, self assessment, formative assessment, peer assessment and other assessments. The others, in minority, have analyzed the validity of tests, like content validity, face validity, index difficulty

⁶ H. Douglas Brown, *Language Assessment*..... 21.

and item discrimination. However, there is a study about essay test which was done by Ita Faradillah entitled “An Analysis of Essay Test on English Final Test for Grade Eleven Students of SMAN 1 Lamongan” that only analyzed the content validity, item discrimination and index difficulty of essay test.⁷ The researcher did not find any studies about reliability, especially intra-rater reliability (internal rater consistency) in scoring essay test which has taken a crucial part like the explanation above. Therefore, the researcher wants to examine this topic thoroughly.

Al-Amin Islamic Boarding School is the pioneer of bilingual Islamic boarding school in Mojokerto. It means that this school becomes a model for other new bilingual schools. As bilingual school, the students have to speak in English or Arabic for their daily conversation. In addition, they get an additional language class in the evening to deepen their language skills, one of them is writing skill.

Absolutely language teachers take important part here. They have to master the language very well in teaching and testing the students. Master of language here means that English teacher must have a college degree and national certificate in language. It can be said that they are professional enough to be good

⁷ 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)

teacher admitted by a college that formed in certificate and have good quality of validity and reliability to be a language tester. As the limitation of the language teachers, a teacher is in charge only for a class. They teach, test and grade their students by them selves, especially in examining students' writing skill, such as an essay test. Therefore, it is very important to know whether the English teachers' intra-rater reliability is consistent or not in scoring essay test to understand how objective they are in assessing subjective test.

B. Problem of the Study

Based on the background of the problem, the research question of study can be formed as follows:

- Is the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto in scoring essay test consistent?

C. Objective of the Study

In this study, the researcher wants to know whether the intra-rater reliability of English teachers was consistent or not in scoring essay test.

D. Hypothesis of the Study

This study has a probability; the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto is consistent in scoring essay test.

E. Significance of the Study

The finding of this study is expected to gain awareness of English teachers about the important of intra-rater reliability consistency in writing assessment, especially essay test. They need to know whether their-self were consistent or not in evaluating their students' writing skill. This awareness will make them become more objective in assessing subjective test. Therefore, English teachers will be more careful in scoring essay test in order that the students get their "real" score and feedback properly.

F. Scope and Limitation of the Study

The scope of this study is to determine whether the intra-rater of English teachers in scoring daily essay test of grade eleven Al Amin Islamic Boarding Senior High School Mojokerto consistent or not. This research will not examine their reliability by comparing each teachers' score (inter-rater reliability) or any other kinds of reliability. Therefore, it will be conducted only in measuring the English teachers' level of intra-rater reliability.

The second limitation of this research is that the researcher conducted the research only once. Once means the research only analyzed their consistency from two scoring (pre- and post-scoring), not repeatedly. As the subjective nature

of this assessment, it has possibility that the result will be changed in different time.

G. Definition of Key Terms

1. Consistency

According to Oxford Dictionary, Consistency is the quality of being consistent. Consistent means behaving the same way, having the same opinions, standards, etc.⁸ In short, consistency is the quality of having the same opinion of something.

In this study, consistency here means the investigation of English language teachers' consistent quality in scoring essay test by obtaining their intra-rater reliability level in statistical analysis of pre- and post-scoring.

2. Intra-rater Reliability

Intra-rater is one of kind of rater reliability which is typically estimated by getting two sets of scores produced by the same raters for the same group of students (say a rater scores one groups' set of compositions on two successive occasions about two weeks apart), and calculating a correlation coefficient between those two sets of scores.⁹

⁸ "Consistency" and "Consistent", *Oxford Dictionary*, 4th Ed. (UK: Oxford University Press, 2008), 91.

⁹ James Dean Brown, *Testing in Language Program* 187.

Intra-rater reliability in this study is to determine English teachers' self-consistency in scoring essay test. Cronbach alpha coefficient is the most appropriate formula to establish this reliability. The result of average has to be above 0.7 in order to be considered that they are internally consistent and reliable. In addition, there will be reliability level for detail information of each rater result.

3. Intra-rater Reliability Consistency

In this research, the intra-rater reliability consistency means the English teachers' self-consistency as raters of essay test. Their own score product in pre- and post-scoring will be examined to decide whether they are consistent or not. As this study investigates about the consistency of intra-rater reliability, the calculation uses intraclass reliability estimation in SPSS 23 to analyze scores produced by raters.

4. Scoring Essay Test

The definition of scoring itself is gaining marks in a test or exams.¹⁰

Essay test is one or more essay questions administered to a group of students under standard conditions for the primary purpose of collecting evaluation data.¹¹ In this research, scoring essay test means grading students' essay as

¹⁰ "Score", *Oxford Dictionary*, 4th Ed. 393.

¹¹ W. E. Coffman, *Essay Examination*. In R.L. Thorndike (Ed.) *Educational Measurement* 2nd ed pp. 271-302 (Washington D.C.: American Council on Education, 1971), 271.

