CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter orderly presents theoretical foundation: definition of rating scale, writing, and assessment. Then, it is continued by the explanation of rating scale in writing assessment, explanation of EBB scale, and EBB scale as a rating scale descriptor in writing assessment. Next is about previous studies, which have the relation of the research.

A. Theoretical Background

1. Rating Scale

In language testing, Namara states that a rating scale is a series of ascending descriptions of salient features of performance at each language level. A language performance can be assessed by examining either the whole impression of the performance or the performance according to different criteria. In this regard, there are two types of rating scales: holistic scales which describe learners' performances as a whole (e.g. the American Council for the Teaching of Foreign Languages scale); and analytic scales which consist of a number of criteria referring to particular aspects of performance such as grammar, fluency and content (e.g. the Test of Spoken English).

2. Writing

Writing is a fundamental aspect of academic literacy and communicative competence in the current educated world.¹ According to Hammad, writing is a form of language outcomes as a real form of language input. In writing, to produce an essay, the authors should have an idea which written with a correct grammatical, mechanical writing, content and communicative language.

Furthermore, writing as a thinking process which involves generating ideas, composing these ideas in sentences and paragraphs, and finally revising the ideas and paragraphs composed. Good writing also requires knowledge of grammatical rules, lexical devices, and logical ties.² Moreover, White and Arndt define 'writing' as "a form of problem-solving which involves such process as generating ideas, discovering a voice with which to write, planning, goal setting, monitoring and evaluating what is going to be written, and searching with language with which to express exact meanings".

3. Assessment

Traditionally, the word "assessment" has referred to the way teachers assign letter grades on tests and quizzes. Assessment has also been used as a way to discuss teaching effectiveness. However, assessment is now

¹ Nadia Behizadeh and George Engelhard, "Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States", *Assessing Writing*, vol. 16, no. 3 (2011), p. 190, accessed 27 Jul 2016.

² Hammad Enas Abdullah, "Palestinian University Students' Problems with EFL Essay Writing in an Instructional Settin", *Journal of Second and Multiple Language Acquisition – JSMULA*, vol. 2, no. 1-21 (2014), p. 2, accessed 31 May 2016.

taking on a new meaning. It should be a "dynamic process that continuously yields information about student progress toward the achievement of learning goals".³ William cited in cornard states that in order for assessment to be considered authentic, it must focus on whether or not students can apply their learning to the appropriate situations.⁴

Assessment used by the teacher to measure student ability in final lesson. The teacher gives assessment to student as a product for the student's input. The assessments are appropriate with students' creativity in applying the lesson. Teacher's goals visible with assessment criteria gives to student, it must success or not student utilize their knowledge of the lesson to get the perfect assessment.

4. Rating Scale in Writing Assessment

The conceptualization of rating scale as part of the test construct has opened a new horizon on examining different aspects of rating scales functioning in performance assessment. Writing assessment as a kind of performance assessment is affected by the quality of the rating scale used. To a large extent, the common thread of arguments on rating scales converges on the important issue of construct validity. Weigle summarizes McNamara, identifying the importance of rating scale in the quality of

³ Garfield, G.B., "Beyond testing and grading: using assessment to improve student learning", *Journal of Statistics Education*, 2(1), 1994. http://www.amstat.org/publications/jse/v2n1/garfield.html

⁴ Nicole Williams, *Reflective journal writing as an alternative assessment* (2008), p. 2,

http://205.133.226.154/Files/pdf/Education/JTIR/VolumeIII/williams.pdf, accessed 31 May 2016.

assessment, on the centrality of the rating scale to the valid measurement of the writing construct:

The scale that is used in assessing performance tasks such as writing tests represents, implicitly or explicitly, the theoretical basis upon which the test is founded; that is, it embodies the test(or scale) developer's notion of what skills or abilities are being measured by the test. For this reason, the development of a scale (or set of scales) and the descriptors for each scale level are of critical importance for the validity of the assessment.⁵

5. Empirically derived, Binary-choice, Boundary definition (EBB) scale

EBB scale is a rating scale descriptors used to assessing with a specific purpose, such as the assessment of productive skills. An EBB has to be developed for each task type in a speaking or writing test.⁶ Fulcher states in his book that in EBB development it is essential to have samples of language (writing or speaking) generated from specific language tasks, and a set of expert judges who will the make decisions about the comparative merit of sets of samples.⁷

The empirically derived, binary-choice, boundary definition scales (EBBs), what distinguishes this method is that the scale – and hence the cognitive process that raters must follow – is set forth as a series of repeated and branching binary decisions. EBBs are constructed by rank ordering performances on test tasks and then identifying key features that

⁵ Batoul Ghanbari, Hossein Barati, and Ahmad Moinzadeh, "Problematizing Rating Scales in EFL Academic Writing Assessment: Voices from Iranian Context", *English Language Teaching*, vol. 5, no. 8 (2012), p. 77, http://www.ccsenet.org/journal/index.php/elt/article/view/18617, accessed 31 May 2016.

⁶ Glenn Fulcher, *Practical language testing* (London: Hodder Education, 2010), p. 212.

⁷ Glenn Fulcher, *Practical language testing*, ...p. 211.

judges use to separate the performances into adjacent levels. EBBs represent an innovation in the logic of how raters judge performance with reference to performance data in specific contexts of language use. EBBs may not contain the rich description of the previous method, but they are relatively easy to use in real-time rating, and do not place a heavy burden on the memory of the raters.⁸

6. EBB scale as a Rating Scale Description in Writing Assessment

EBB scales develop for productive skills that construct based on student ability in the same level. The step to develop a rubric using EBB scale is; According to Upshur and Turner there are six steps to develop EBB scale:

- a. Eight student performances were selected from the set to be rated. These should represent approximately the full range of ability in the total set.
- Each of the six members of the research team individually divided the set of eight performances into the four better and four poorer. This was done impressionistically.
- c. The team discussed their dichotomous rankings and reconciled any differences. They then formulated the simplest criterial question that would allow them to classify performances as 'upper-half' or 'lowerhalf' according to the attribute that they were rating. Similarly, to construct an eight-point scale, three levels of questions would be needed; the rater would ask three questions to score any writing

⁸ G. Fulcher, F. Davidson, and J. Kemp, "Effective rating scale development for speaking tests: Performance decision trees", *Language Testing*, vol. 28, no. 1 (2011), p. 9, accessed 31 May 2016.

sample. With a fourth level of questions, a sixteen-point scale can be made.

- d. Working with the four upper-half performances, the team members individually rated each of them as '6', '5', or '4'. The procedure requires that at least one sample should be rated as '6'; at least two numerical ratings must be used. Therefore, at least two of the four same rating. samples receive the This scoring was done impressionistically. A six-point scale was chosen for two reasons. First, a scale with an even number of categories allows for a binary split into two equal halves by means of the first criterial question. Secondly, six categories can be readily handled by judges, and provide relatively high reliability of ratings.⁹
- e. Rankings were discussed and reconciled. Simple criteria questions were formulated, first to distinguish level 6 performances from level 4 and 5 performances, and then level 5 performances from level 4 performances.
- f. Steps 4 and 5 were repeated for the lower-half performances.¹⁰

B. Previous Study

In this part, the researcher reviewed some previous studies related to this research:

⁹ Finn, R. H., "Effects of some variations in rating scale characteristics on the means and reliabilities of ratings". 1972. *Educational and Psychological Measurement 32: 255-65*.
¹⁰ John A. Upshur and Carolyn E. Turner, "Constructing rating scales for second language tests", *ELT journal*, vol. 49, no. 1 (1995), p. 7, accessed 31 May 2016.

- 1. First, there are also some researches related to using rating scale for assessment in productive skills. Harai and Koizumi were done the study. They were analyzing validity and reliability between two rating scale, those are; EBB scale that compared with an analytic scale. The research was conducted to know the rating scale more valid and reliable to assessing productive skills. This research using Story Retelling Speaking Test (SRST) to verify two rating scales. The finding shows that EBB scale more valid and reliable for SRST, whereas analytic scale more practically. The EBB scale was slightly superior in reliability and validity, whereas the analytic scale excelled in practicality. However, the results helped us find points for revision in the scales. First, the descriptors of EBB Grammar & Vocabulary criterion should be modified. Second, the Communicative Efficiency and Content criteria of the EBB scale can be combined to enhance its practicality. Third, the current EBB binary format might be changed into one similar to the analytic scale. Since both scales have strengths, combining the good aspects of these scales may enable us to create a better scale in the future.¹¹
- 2. Second is the research was done by Mu-shuan Chou, entitled "Teacher Interpretation of Test Scores and Feedback to Students in EFL Classrooms: A Comparison of Two Rating Methods". The research was conducted to know interpreted teacher in scoring speaking and student's benefit from feedback of two rating scale descriptor. This research using

¹¹ Harai Akiyo, and Koizumi Rie, "Validation of the EBB scale: A Case of the Story Retelling Speaking Test", p. 15.

rating scale and rating checklist that focused on the oral test in role play and simulation task to compare teachers interpret and student's feedback. The rating scale descriptor are adopt for this research is the empirically derived, binary-choice, boundary definition (EBB) scale and Performance Decision Tree' (PDT) scale. The finding shows that rating scale more detail in description and help student more focus than rating checklist. The criteria in the performance data-driven rating scale and the rating checklist were the same; the fifteen teachers reported differing interpretations of student performance on the role-play/simulation tasks. Eleven out of fifteen (73.3%) considered that the rating scale, with its more detailed descriptors, offered more comprehensive information of students' speaking ability and skills than the rating checklist. Seven said that the scale was effective in terms of helping them decide whether their students were able to use what they had learned in class and whether they used it correctly or not. Four teachers reported that they tended to focus more on what had *not* been done by the students in the task than pay attention to what had been done, so the rating scale with detailed descriptors helped them focus on positive aspects of student performance.¹²

 Third, is "Problematizing Rating Scale in EFL Academic Writing Assessment: Voice from Iranian Context" by Batoul Ghanbari, Hossein Barati & Ahmad Moinzadeh. The research was conducted to know

¹² Mu-hsuan Chou, "Teacher Interpretation of Test Scores and Feedback to Students in EFL Classrooms: A Comparison of Two Rating Methods", *Higher Education Studies*, vol. 3, no. 2 (2013), p. 90, http://www.ccsenet.org/journal/index.php/hes/article/view/24808, accessed 31 May 2016.

effectiveness using local rating scale. This research was compared the assessment used traditional approach and rating scale. The research was done with some questioner given to teacher or test-taker to know how far the teacher used local rating scale. Majority of the raters in this study believed that native scales have to be appropriated in the context before application. In their ideas, unmediated application of native rating scales would surface a hidden conflict between the assumptions behind these scales on the one hand and the realities of the local context on the other hand. McNamara states that strongly questions the validity of rating scales and by tracking the origin of scale tradition in the FSI test in the 1950s shows how successive rating scales developed over the last four decades have been heavily influenced by the assumptions, and even the wording of the original work, and rare empirical validation has been done.¹³

4.

5. Fourth, the other study was done by Batoul Ghanbari, Hossein Barati and Ahmad Moinzadeh, entitled "Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny". The research is analyzing the construct of rating scale for writing assessment. Such as the recent study, "Problematizing Rating Scale in EFL Academic Writing Assessment: Voice from Iranian Context" show that the rating scale was construct with two focuses, the socio-cognitive and critical argument, deficiencies of the present practice of adopting rating scales are revealed and consequently it

¹³ Ghanbari, Barati, and Moinzadeh, "Problematizing Rating Scales in EFL Academic Writing Assessment", p. 85.

- is discussed how assessment circles in native countries by setting rating standards control and dominate the whole process of writing assessment.¹⁴ The findings shows that a local rating scale has a more valid outcomes, but test takers should always develop it to be more appropriate to student's writing assessment in Iran. This article reminds that in addition to the long-term obligation of continually examining and testing the evaluation procedures and the assumptions that underlie them, a local rating scale as it takes into account the particularities of each assessment context would lead to more valid outcomes.¹⁵
- 6. Fifth is "Benefits from Using Continuous Rating Scales in Online Survey Research" by Horst Treiblmaier and Peter Filzmozer. The research was conducted to know the effectiveness using continuous rating scales in online survey. This research is compared manual measurement and computers administer surveys. The major problem with such scales in the past was the inaccurate and tedious data collection process, since with pencil-and-paper surveys it was necessary to manually measure the respondent's answer on a sheet of paper. The usage of computer-administered surveys render such problems obsolete, since measurement and data collection can be done without any loss of precision.¹⁶ This

¹⁴ Batoul Ghanbari, Hossein Barati, and Ahmad Moinzadeh, "Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny", *Language Testing in Asia*, vol. 2, no. 1 (2012), p. 83, accessed 31 May 2016.

¹⁵ Batoul Ghanbari, Hossein Barati, and Ahmad Moinzadeh, "Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny",...p. 97.

¹⁶ H. Treiblmaier and P. Filzmoser, 2009. "Benefits from Using Continuous Rating Scale", *Institut f. Statistik u. Wahrscheinlichkeitstheorie*, vol. 107, no. 8-10, p. 19.

research applies the robust MCD (Minimum Covariance Determinant) estimator to a data set which consists of variables with a data range from 1 to 100. The results show that outliers (which can occur in the form of "nonsense" data or noise in any survey) severely affect the correlation of the variables. In a next step, we illustrate that the application of robust factor analysis, which can be applied only with the 100-point scale, leads to more pronounced results and a higher cumulative explained variance. Given that factor analytic procedures are part of covariance based structural equation modeling, which is frequently used to test theories in IS research, our findings also bear huge significance for such advanced techniques.¹⁷

7. Sixth is "Effective Rating Scale Development for Speaking Tests: Performance Decision Trees" by Glenn Fulcher, Fred Davidson, and Jenny Kemp. This research was done to develop a Performance Decision Trees (PDT) scale to be easier to use for speaking test. Performance Decision Trees are more flexible and do not assume a linear, unidimensional, reified view of how second language learners communicate. They are also *pragamatic*, focusing as they do upon observable action and performance, while attempting to relate actual performance to communicative competence.¹⁸ This research show that the rating scale of speaking test be easier to apply but it still valid and reliable. A major part of a validity

 ¹⁷ H. Treiblmaier and P. Filzmoser, 2009. "Benefits from Using Continuous Rating Scale", ...p. 19
 ¹⁸ G. Fulcher, F. Davidson, and J. Kemp, "Effective rating scale development for speaking tests: Performance decision trees", *Language Testing*, vol. 28, no. 1 (2011), p. 23, accessed 1 Jun 2016.

claim for a PDT would rest upon the comprehensiveness of the description upon which it was generated, and the relevance of the assessment categories to current theories of 'successful interaction' within a particular context. As such, other PDTs must be developed through a careful analysis of communication in context, and a theoretical description of the constructs that underlie successful interaction, in order to generate context sensitive assessment categories.¹⁹

8. Seventh is "Pengembangan Instrument Penilaian Membaca Kelas VII SMP" by Nila Maulan, Imam Agus Basuki, and Bustanul Arifin. This research was done to develop an instrument for reading assessment. The product of this research was an instrument used to assess reading dictionary, reading fast, reading ceremony text, retelling story, and giving comment on narrative text. There were two instrument for assessing reading dictionary; those are assignment of reading dictionary and rubric for assessment. Instrument for assessing fast reading were subjective text and answer clue. Instrument for reading ceremony text were reading ceremony text and rubric assessment. Instrument for assessing retelling story were retell the story and rubric assessment. And instrument for assessment. According to evaluator expert, practice expert and students,

¹⁹ G. Fulcher, F. Davidson, and J. Kemp, "Effective rating scale development for speaking tests: Performance decision trees", ...p. 25.

the researcher concluded that based on validity, reliability, and practices the product appropriate to be implemented.²⁰

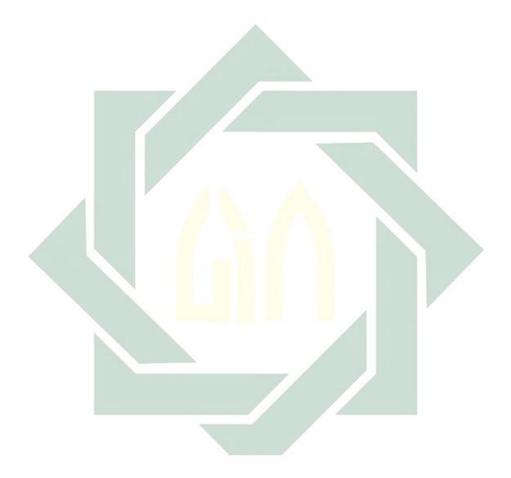
9. Eighth is "Pengembangan Rubrik Penilain Portofolio Proses Sains Siswa pada Materi Ekosistem di SMPN 1 Wedarijaksa-Pati" by Vera Widyaningsih. This research was done to develop rubric for sains assessment. Based on the research, the writer concluded that the teacher did not use rubric for each aspect on the lesson. Teacher used rubric for assess cognitive aspect in *LKS* only. The researcher develop rubric for assessment through identify process, construct rubric, and implementation the rubric. Based on the result, students mean had not differences between three classes. Teacher and students responses were positive for rubric assessing sains. Rubric for portfolio assignment for ecosystem material to be appropriate for assessment.²¹

This research is different from all of those researches because in this research, the researcher analyze the development and implementation of empirically derived, binary choice, boundary definition scale in senior high school. The procedure to develop rating scale descriptors of empirically derived, binary choice, boundary definition scale is first give the task to a group of students drawn from the target population. Take the resulting language samples and ask the group of experts to divide them into two groups

²⁰ N. Maulani, I. Agus Basuki, and B. Arifin, "Pengembangan Instrument Penilaian Membaca Kelas VII SMP", *Universitas Negeri Malang*. 2012

²¹ V. Widyaningsih., Undergraduate thesis; "Pengembangan Rubrik Penilaian Portofolio Proses Sains pada Materi Ekosistem di SMPN 1 Wedarijaksa-Pati". (Universitas Negeri Semarang, 2013)

– the 'better' and the 'weaker' performances.²² The researcher uses descriptive qualitative research.



²² Glen Fulcher, *Practical language testing*, ...p. 211.