Conference Paper

Severity-Leniency in Writing Assessment and Its Causal Factors

Nur Azizah[1], Muchlas Suseno[2], and Bahrul Hayat[3]

[1]Educational Research and Evaluation Program, State University of Jakarta (UNJ), Rawamangun, Jakarta
[2]Department of English Language and Literature, State University of Jakarta (UNJ), Rawamangun, Jakarta
[3]Departement of Psychology, UIN Syarif Hidayatullah Jakarta, Jakarta

Abstract

Objectivity in a writing assessment is a highly important matter because it determines the validity and reliability of the writing assessment itself. However, subjectivity in the writing assessment is inevitable. Rater subjectivity in the assessment can lead to assessment difference or variability (severity-leniency) eventually reducing the validity and reliability of the assessment. Therefore, the concept of rater variability, severity-leniency, and its causal factors are required to perceive so that efforts to minimize assessment variability, for instance, severity-leniency, can be implemented. The research aimed at providing a brief description of rater variability, namely severity-leniency, and the causal factors, namely rater's background, criteria, method, and assessment scale. The method used in this research was literature study which source was originated from the articles of scientific journals and books related to the research topic, namely rater variability (severity-leniency) and its causal factors. The result of the discussion in this paper showed that the selection of experienced rater, coherent and cohesive assessment criteria, method appropriate to the aim of the research, and brief assessment scale were considered to be able to minimize the variability (severity-leniency) of the rater.

Keywords: leniency, rubric, severity, variability, writing assessment

Corresponding Author:
Nur Azizah
nurazizah_pep17s3@mahasiswa.unj.ac.id

Published: 11 November 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the IC-HEDS 2019 Conference Committee.

1. Introduction

Writing is a complex process necessary to expand the learning, thinking, and communicating with other people [1]. Even in Indonesia, producing academic writing and publishing are the necessity as the graduation requirement to fulfill by bachelor, master, and doctoral students [2]. Based on that policy, it is considered that writing is a highly necessary matter because it influences the future and career of many people.

However, there are problems related to the writing activity, particularly in a writing assessment. The success of the writing assessment is determined by rater [3]. The assessment involving rater generally had disadvantages, namely subjectivity [4].

The rater subjectivity in the research would make assessment difference or variability [5]. Rater variability occurred in several aspects, namely interpretation of assessment criteria, understanding in the usage of assessment scale, assessment consistency, and severity-leniency [6]. Rater variability, variation, or difference often occur in the assessment and what interested the researcher most is severity-leniency.

Therefore, this research aims at describing the factors contributing to rater variability, namely severity-leniency, and its causal factors, namely rater's background, assessment criteria, assessment method, and assessment scale. This study is necessary because by arranging and exploring the matters related to the writing assessment, it is expected that it possibly identifies one's writing development [7].

2. Methods

This paper aims to explain the concept of severity leniency and the factors that cause it. In achieving this goal, a literature study approached is used by using references in the form of scientific journal articles and books. This literature review focuses on research conducted from the 1980s to 2019s. To collect related journal articles are focused on Google Scholar and international journal provider sites, namely the Directory of Open Access Journal (DOAJ), academia.edu, research gate, and science direct. The keywords used *severity, leniency, rater, severity rater, leniency rater, writing assessment, writing assessment method, rubric, holistic rubric, analytical rubric, grading scale*, and *rubric scale*. More than 100 journal articles, 20 books, 10 papers, and 5 reports were collected. The references collected which most relevant to the topic, namely about severity-rater efficiency and the factors that cause it. Finally, there are 24 articles, 9 books, 4 papers, and 1 report collected as samples to be analyzed. The subjects analyzed from the articles which the concept severity-leniency, the factors causing severity leniency, and results of the research. The analysis did by comparing the concepts and results of research, related to severity-leniency, factors that cause it, and then synthesizing it.

3. Results

3.1. Rater Variability (Severity-Leniency)

Variability, variation, or difference of assessment result was an inevitable factor in the assessment involving rater because of the subjectivity [8]. Linacre [9] stated that rater variation or difference in the assessment was an inevitable matter in the writing

assessment. The statements of the experts could be proven by the research related to the variability of writing assessment involving rater. One of them was the research conducted by Schaefer [10].

In his research, Schaefer [11] involved 40 native speakers of English to evaluate 40 essays written by the students of English program in Japan. Each rater evaluated 40 essays by using analytical assessment scale with five categories (content, organization, style and expression quality, language usage, and fluency). The result of his study using Many Facets Rasch Measurement (MFRM) showed that content or organization category had quite a tight score (severity) and the language usage category had quite a loose score (leniency). He also found that the raters evaluated the great-skilled writers more tightly (severe) than those with the low skill (lenient).

Severity-leniency assessment referred to the rater's tendency to consistently give a lower or higher score than the score it is supposed to be [12]. This definition is in line with the opinion of other experts stating that severity-leniency is the rater's tendency in conducting an assessment which advantages or harms the people being evaluated [13]. The definition stated by Engelhard [14] is that the rater's tendency in evaluating higher or lower than how it was supposed to be, is seen to be more appropriate. It was because the measurement of higher and lower stated by Engelhard [15] was clearer and easier to be measured than the assessment term "advantage or disadvantage" as stated by Knoch [16].

3.2. The Causal Factors of Rater Variability

The background, mother tongue, experience, and training of the rater were stated as the factors influencing the writing assessment. The factors were believed to be able to influence the accurateness, accuracy, and justness in the writing assessment. The researcher found that the factors related to the rater including the rater's background, assessment method used, assessment criteria, professional experience, and error tolerance [17].

Shohamy [18] studied several factors which could influence assessment, namely the rater's background, training, and assessment scale. According to the experts' opinions and library research, the factors influencing rater variability could be concluded into four categories, namely rater's background (including training and experience), assessment criteria, assessment method, and assessment scale.

3.3. The Background of the Rater

Rater's background was the factor believed to be able to cause variability or differences between raters related to severity or tightness in assessment [19]. It was in line with the opinion of Barret [20] stating that the background factor was one of the most important matters because, besides assessment criteria, it was what caused rater variability.

Rater's background included personal background, professional background, and evaluating experience [21]. Personal background generally is determined by age, gender, education, etc. Professional background or rater's occupation can influence the rater's reliability. Teacher or tutor, for instance, can be influenced by the instructional aims and particular components in the writing assessment. On the other hand, novice rater could evaluate based on their personal view of the writing. Cumming [22] stated that the rater's experience referred to how long the rater had been evaluating or the number of the assessment completed by the rater. Several studies about experienced raters and inexperienced raters showed different interrater consistency [23].

There was much research investigated the difference between experienced/expert rater and novice rater to bring the problem of rater's background [24]. The research of Cumming [25] showed that compared to the novice rater, expert raters tended to use some wider criteria. Wolfe [26] showed that experienced raters tended to focus on more general (holistic) assessment criteria than the novice raters did. Song and Caruso [27] also found that raters having more experience tended to evaluate more softly than those having less experience did when using holistic assessment.

3.4. Assessment Criteria

Popham [28] argued that criteria are a tool used in assessing student's competence in a certain field. Brookhart [29] asserted that good assessment criteria are the existence of coherency inter-assessment criteria and the existence of level or stage in assessment criteria. Criteria in writing assessment column should be coherent which means that inter-criteria should be compatible or there is no contradictory statement in criteria. In writing assessment criteria, there is also a level or stage to assess performance, work, or evaluated writing activity done individually. As assessment guideline which contains assessment criteria, the column is explained as follows. Mertler [30] argued that assessment criteria should have been settled beforehand to make compatibility with the determined assessment goal. Thus, assessment criteria help an individual to

understand why they get a certain score in one of assessment and what they should be done to increase their performance for the next assessment.

Stevens and Levi [31] explained that assessment criteria as part of specific feedback. It is mentioned so because this part enables the assessor in giving specific feedback to an evaluated individual. The good assessment criteria can provide detail feedback about parts of the assignment and how good or bad performance which is done. In assessment criteria, there are details of task assessment components which can be known by other individuals. Thus, an evaluated individual can evaluate themselves in what part their weakness and strength performance is. In this assessment criteria, an evaluated individual also gets assessment transparency since the obtained score can be investigated.

3.5. Assessment Method

There are two types of assessment methods in writing, namely holistic and analytical. The holistic method is used to assess the entire processes or products without assessing the components of the products or processes separately [32]. Conversely, the analytical method is used to conduct scoring separately. The product or performance is assessed at first. Then, the assessment score is summed to obtain a total score.

A holistic method focuses on the whole assessment process regardless of the parts. In holistic scoring, the focus of the assessment is directed at the holistic or overall writing performance rather than on certain aspects of the essay such as content, organization, grammar, punctuation, and so on. It causes such scoring is not acceptable to measure the specific competency in writing skills. However, this assessment method is considered more practical because the assessor does not need to read it repeatedly to give a rating. It makes the holistic assessment process faster and more comprehensive.

This assessment method usually contains a description of the highest level of performance to describe in the comment column for each criterion to describe the achievement of individual performance at that level. Nitko [33] stated that the application of the holistic method is more appropriate if the given task does not have a definite correct answer. The focus of holistic method application lies on the overall quality, skills or content's understanding specifically, and other skills which involve assessment at a unidimensional level [34].

The assessment process by the holistic method is relatively faster than that by the analytical method [35]. It is because the assessor does not have to repeatedly conduct the assessment [36]. Since the key is the overall performance assessment, a holistic

method is used for summative performance assessment. Expected feedback of the rubric assessment in this type is limited or not excessive. In a holistic assessment, assessment criteria are possible in the form of a combination or a combination of criteria placed on a single descriptive scale [37]. The assessment by a holistic method supports a broader assessment of the quality of the process or product. The assessment format by using the holistic method according to Mertler [38] is shown in Table 1.

Table 1: Holistic method assessment

Score	Description
5	Showing a complete understanding of the problem. There are all requirements of the task in the answer.
4	Showing a sufficient understanding of the problem. There are all requirements of the task in the answer
3	Showing a partial understanding of the problem. There are most of the requirements of the task in the answer
2	Showing a little understanding of the problem. There are most of the requirements of the task not in the answer
1	Showing no understanding of the problem.
0	No answers / no effort

Arter and McTighe [39] suggested that the analytic method assessment assesses something based on the essential traits or dimension assessed separately. In a meta-analysis from 75 studies about the implementation of the analytic method, Jonsson and Svingby [40] reported that analytic method assessment can be an effective tool. Analytic-method assessment allows us to separately evaluate on each component, factor, or its part. Each criterion in analytic method assessment is assessed based on a different scale [41]. Analytical-method assessment is employed when the needed responses of the individual are focused [42]. For the procedure of the assessment, assessment result by analytic method initially is in the form of scores, and the total score is eventually summed up.

Analytic method assessment focuses on the scoring of components assessed by calculating the mistakes in detail. The total score is the merger of the scores from each component. The advantage of this scoring technique is that the teacher as the rater is allowed to assess all elements supporting students' writing skills in more detailed. For the students, this scoring method helps them understand the elements that must be concerned in a text. Moreover, the students are allowed to evaluate whether their writing is good or not by using the assessment criteria. The disadvantage is the difficulty in the quantification of assessment result on each component. It requires further thought so that the assessment can be effective, reliable, and objective.

As mentioned previously, the use of analytic method assessment causes the process of the assessment slower, especially because the assessment of some skills or characteristics are individually different. Consequently, the rater must check the product or process more than once. Both construction and its implementation can be too time-consuming. Individual work should be checked in a different time for each particular performance task or assessment criteria [43]. Moreover, the analytic method assessment possibly causes the overlap among the determined criteria.

However, the advantage of using the analytic method is quite significant, that is the special feedback related to individual performance on each assessment criterion. This feedback is not available in holistic method assessment [44]. Therefore, holistic method assessment allows the rater and the assessed ones to draw the strengths and weaknesses of individual performance assessed [45]. The assessment format of analytic-method according to Stevens and Levi [46] is shown in Table 2.

Table 2: Analytic method assessment

	Beginning	Developing	Accomplished	Exemplary	Score
Criteria #1	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #2	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #3	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #4	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	

3.6. Assessment Scale

The scale is the achievement level in the form of a score. The scale represents how well or badly the task is done [47]. One of the scales commonly used is arranged by Huba and Freed [48]; such as *sophisticated, competent, partly competent, not yet competent; exemplary, proficient, marginal, unacceptable; proficient, medium-high, middle, novice*;

and *distinguished, capable, intermediate, beginner*. Jacobs [49] arranges the assessment scale by using quality gradations from *very poor to excellent,* while the assessment scale designed by Hamp-Lyons [50] and Andrade [51] uses quality gradations from *unsatisfactory, fair,* to *good.* Assessment scale employed in *San Diego Unified School District* uses four quality gradations from *below basic, basic, proficient,* up to *advanced.* The difference is also shown from the quality score made for each assessment category. For organization, it is given maximum quality score 20, elaboration of idea/content is given maximum quality score 30, grammar is given maximum quality score 25, vocabulary is given maximum quality score 20, and mechanism is given maximum quality score 5. Thus the total score that can be obtained by students is 100.

Different from Jacobs, an assessment scale designed by Hamp-Lyons and Andrade also assessment scaled employed in *San Diego Unified School District* do not use the quality score as the assessment scale of Jacob does. Both types of rubric use numbers to represent the quality obtained by students. An assessment scale of Hamp-Lyons and Andrade, the quality gradation of *unsatisfactory* is given quality score 1*, fair* is given quality score 2, and *good* is given quality score 3. Thus, the maximum score that can be obtained by students is 18 points with its minimum score of 6 points. The same as the assessment scale by Hamp-Lyons and Andrade, assessment scale employed in *San Diego Unified School District* also uses numbers to explain the quality obtained by students. Nevertheless, this rubric matches the number of quality gradations used as many as the quality score used, namely 4 points: 1 for *below basic,* 2 for *basic,* 3 for *proficient,* and 4 for *advanced.*

4. Discussion

It is not that difficult to minimalize the variability (*severity-leniency)* and achieve high inter-rater reliability. It surely requires optimum efforts. An important effort to minimize the severity-leniency is by investigating the background of the rater. It is important since the background of the rater is the factor that potentially causes the inter-rater variability or difference related to *the severity* of their strictness in assessing [52]. Another effort is by assessor training before assessing. Assessor training indeed cannot eliminate rater variability, but the rater can be more consistent and improve the reliability of the assessment [53].

The decision related to the use of holistic or analytic method assessment for writing an assessment to minimalize the severity-leniency of the rater has some possible implications. The most significant one is the consideration of the goal of assessment and how

the assessment result is employed. If the assessment is general and summative, holistic-method assessment is more suitable. Vice versa, if the goal is formative feedback, the suitable assessment is the analytic method. The time requirements, the nature of the task, and the specific performance criteria observed must also be considered.

In terms of the selection of the rating scale to minimalize severity-leniency of the rater, there is no definite formula to determine the number of levels or length of the scale used in the assessment scale. Generally, experts use three-level-scale up to five-level-scale. The more the number of levels is, the more difficult it is to distinguish and explain precisely why an assessment is included in that level. Most experts consider the assessment of three levels as the optimal assessment level of assessment scale [54].

5. Conclusion

To minimalize the variability of the rater in writing an assessment, considering the causal factors of severity-leniency is a must. Those are the rater's background, the criteria, the method, and the assessment scale. The efforts to choose the experienced rater, coherent and non-overlapping assessment criteria, a method suitable to assessment goal and short assessment scale are believed to be able to minimize the variability (severity-leniency).

Acknowledgement

The authors would like to thank their colleague for their contribution and support to the research. They are also thankful to all the reviewers who gave their valuable inputs to the manuscript and helped in completing the paper.

Conflict of Interest

The authors have no conflict of interest to declare.

References

[1] Dunsmuir, S., & Clifford, V. (2003). Children's Writing and the use of ICT. *Educational Psychology in Practice*, vol.19, pp. 171–187.

[2] Circular Letter of Dirjen Dikti No. 152/E/T/2012

[3] Crooks, Terry & Kane, Michael & Cohen, Allan. (1996). Threats to the Valid Use of Assessments. *Assessment in Education*: *Principles, Policy & Practice*. vol. 3, pp. 265-286.

[4] Messics, S. (May 1996). *Validity and wahsback in language testing* (Report No. RR-97-17, pp. 241-256). Educational Testing Service.

[5] Schaefer, B., Fricke, S., Szczerbinski, M., Fox-Boyer, A. V., Stackhouse, J., & Wells, B. (2009). Development of a test battery for assessing phonological awareness in German-speaking children. *Clinical Linguistics and Phonetics*, vol. 23, pp. 404-430.

[6] Lumley, T. (2005). *Assessing second language writing*: *The rater's perspective*. Frankfurt: Peter Lang.

[7] Herrington, A., & Curtis, M. (2003). Writing development in the college years: By whose definition? *College Composition and Communication*, vol. 55, pp. 69–90.

[8] Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A Many-Facet Rasch Measurement of Differential Rater Severity / Leniency and Teacher Assessment. *JALT Journal*, vol. 34, pp. 79-102

[9] Linacre, J. M. (1989). *Many faceted Rasch measurement*. Chicago: MESA Press. Linacre,

[10] Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A Many-Facet Rasch Measurement of Differential Rater Severity / Leniency and Teacher Assessment. *JALT Journal*, vol. 34, pp. 79-102.

[11] Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency and Teacher Assessment. *JALT Journal*, vol. 34, pp. 79-102.

[12] Engelhard, G. (2012). Examining Rating Quality in Writing Assessment: Rater Agreement, Error, and Accuracy. *Journal of Applied Measurement*, vol. 13, pp. 321-35.

[13] Knoch, U., Read, J., & Randow, J. Von. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing,* vol. 12, pp. 26-43.

[14] Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement,* vol. 31, pp. 93–112.

[15] Engelhard, G. (2012). Examining Rating Quality in Writing Assessment: Rater Agreement, Error, and Accuracy. *Journal of Applied Measurement*, vol. 13, pp. 321-35.

[16] Knoch, U., Read, J., & Randow, J. Von. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing,* vol. 12, pp. 26-43.

[17] Huang, J., & Foote, C. J. (2010). Grading Between the Lines: What Really Impacts Professors ' Holistic Evaluation of ESL Graduate Student Writing? Grading Between the Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing*? Language Assessment Quarterly,* vol. 7, pp. 219–233

[18] Shohamy, E., Aviv, T., Gordon, C. M., Aviv, T., & Kraemer, R. (1992). *The Effect of Raters ' Background and Training on the Reliabilityvof Direct Writing Tests*.The Modern Language Journal, vol. 76, pp. 27-33.

[19] Mcnamara, D. S., & Kintsch, E. (2009). Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning From Text Are Good Texts Always Better? *Cognition and Instruction*, vol. 14, pp. 1-43.

[20] Barrett, Steven. (2001). The impact of training on rater variability. *International Educaion Journal*, vol. 2, pp. 49-58.

[21] Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M.Williamson and B. A. Huot, *Validating holistic scoring for writing assessment. Theoretical and empiricalfounaiztions* (pp. 23765). Cresskill, NJ: Hampton Press, Inc.

[22] Cumming, A. (1990). Language Testing Expertise in Evaluating Second Language Compositions. *Language Testing*, vol. 7, pp. 31-51.

[23] Attali, J., Benaissa, A., Soize, S., Kadziolka, K., Portefaix, C., & Pierot, L. (2014). Follow-up of intracranial aneurysms treated by fl ow diverter: comparison of three-dimensional time-of- flight MR angiography (3D-TOF-MRA) and contrast- enhanced MR angiography (CE-MRA) sequences with digital subtraction angiography as the gold stand. *J NeuroIntervent Surg*, pp. 1–6.

[24] Besterfield-Sacre, M., Gerchak, J., Lyons, M. R., Shuman, L. J., & Wolfe, H. (2004). Scoring concept maps: An integrated rubric for assessing engineering education. *Journal of Engineering Education*, vol. 93, pp. 105–115.

[25] Cumming, A. (1990). Language Testing Expertise in Evaluating Second Language Compositions. *Language Testing*, vol. 7, pp. 31-51.

[26] Wolfe, E. W. (1998). A two-parameter logistic rater model (2PLRM): Detecting rater harshness and centrality. *Paper presented at the annual meeting of the American Educational Research Association*. San Diego, CA.

[27] Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, vol. 5, pp. 163-182.

[28] Popham, W. J. (2010). Instructional sensitivity. In W. J. Popham (Ed.), *Everything school leaders need to know about assessment*. Thousand Oaks, CA: Sage.

[29] Brookhart, S. M. (2001). *Developing Measurement Theory for Classroom Assessment Purposes and U s e s*. ASCD: Virginia, USA.

[30] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research & Evaluation*, vol. 7, pp. 1–10.

[31] Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning.* Sterling, VA; Stylus Publishing.

[32] Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

[33] Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

[34] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research Evaluation*, vol. 7, pp. 1–10.

[35] Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

[36] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research Evaluation*, vol. 7, pp. 1–10.

[37] Dunbar, N. E., Brooks, C. F., & Kubicka-miller, T. (2006). Oral Communication Skills in Higher Education: Using a Performance-Based Evaluation Rubric to Assess Communication Skills 1. Innovate Higher Education, vol. 31, pp. 115–128.

[38] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research Evaluation*, vol. 7, pp. 1–10.

[39] Arter, J. & McTighe, J. (2001). *Scoring Rubrics in the Classroom*. Thousand Oaks, CA: Corwin Press.

[40] Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, vol. *2*, pp. 130–144.

[41] Brookhart, S. M. (2001). *Developing Measurement Theory for Classroom Assessment Purposes and U s e s*. ASCD: Virginia, USA.

[42] Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

[43] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research Evaluation*, vol. 7, pp. 1–10.

[44] Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

[45] Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Mertler, Craig A. *Research Evaluation*, vol. 7, pp. 1–10.

[46] Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning.* Sterling, VA; Stylus Publishing.

[47] Sumintono, B. (2014). Model Rasch untuk Penelitian Sosial Kuantitatif. *Papers of Public Lecture at Department of Statistics, ITS Surabaya, 21 November 2014*, pp. 1–9.

[48] *Huba*, M. E., & *Freed*, J. E. (*2000*). Learner-centered assessment on college campuses: Shifting the focus from teaching to learning. Boston: Allyn and Bacon.

[49] Jacobs, H.L. et al. (1981). *Testing ESL Composition: A Practical Approach.* New York: Newbury House Publishers,

[50] Hamp-Lyons, Liz. (1992). Holistic Writing Assessment of LEP Students. *Proceedings of the National Research Symposium on Limited English Proficient Student Issues* (2nd, Washington, DC, September 4-6, 1991)

[51] Andrade, H. L., Du, Y., & Mycek, K. (2010). Assessment in Education: Principles, Policy & Pctice Rubric – referenced self – assessment and middle school students ' writing. *Assessment in Education*, vol.17, pp.199-214.

[52] McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

[53] Wang, Binhong. (2010). On Rater Agreement and Rater Training. English Language Teaching. *English Language Teaching,* vol. 3, pp. 108-112.

[54] Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning.* Sterling, VA; Stylus Publishing.