RATERS' BIAS, BACKGROUND AND PERCEPTION IN AWARDING SCORE OF WRITING PERFORMANCE

Endah Yulia Rahayu
Postgraduate in ELT of Universitas Negeri Malang, East Java Indonesia
English Department of Teacher Training Faculty
Universitas PGRI Adi buana Surabaya, East Java, Indonesia
Email : indahr_99@yahoo.com

Abstract

Assessing writing performance commits bias due to interaction between raters and criteria because raters can score more consistently or harshly on some criterions. Therefore I explored how the seven raters assessed three essays in order to seek their bias in their rating task, how their background effect (having teaching writing experience & length of teaching writing) their scoring, and how their perception understanding the scoring rubric. The instruments were three essays, analytical writing rubric, questionairres of raters' background and perception. I applied Two-Way Anova, One-Way Anova and Hoyt's Anova to measure the raters' bias, background and perception in awarding score of writing performance. The raters' scoring criteria of Content, Organization and Vocabulay (0.195, 0.511, 0.545 $<$ 0,600) were respectively found bias. Based on the raters' background of having experience of teaching writing, the scoring criteria of Mechanics was bias (0.026 $<$ 0.050). But the length of teaching writing experience did not affect the scoring criteria of Content, Organization, Vocabulary, Language Use and Mechanics, in term of no bias (0.705, 0.663, 0.171, 0.206, 0.090 $\geq$ 0.050). Based on the raters' perception questionnaire, they were familiar with the instrument of writing rubric prior to this reseach and agreed that the rubric help them to discriminate among the different score level. They also considered that the rangefinders in the rubric were usefull tools to asign score, and the writing rubric measured some essential elements for effectively teaching and learning writing. They assumed the rubric could be used as a professional development tool to support teaching and learning writing, and finally they were confident in their ability to score using the rubric.

Keywords: bias, background, perception, writing rubric

INTRODUCTION

In assessing writing perfomance, undeniably raters can commit bias in scoring and turn out to be problematic because they seem to be too emotional and linient over content in students' essays. They could be professional by not letting any bad day affect their grading and they can set a peacefully secured and quit working with their students' written works. However for today's teachers who have many works not only in the class but also before and after their class, it is hard to do. Thus, no assessment is free from bias. Some extraneus factors influecing a student's essay score include: (a) The nature of the particular writing prompt or task posed, (b) the particular rater(s) who judged the student's essay, (c) situation-

specific factors associated with the particular rating occasion, (d) the student's background and interest in the topic or problem presented, and (e) interactions among these different sources. (Sudweeks, Reeve, & Bradshaw, 2005)

Based on validity terms, bias can be seen as *'construct-irrelevant variance that distorts the test results and therefore makes conclusions based on scores less valid'*. In assessment, it is directly related to fairness which conveys "*a skewed and unfair inclination toward one side (group, population) to the detriment of another*" (McNamara, Roever, 2006). If students essays are scored differently on a test or item from their equal ability, a construct-irrelevant variance affects their scores, causing the unidimensional test to become multidimensional. (Saeidi, Yousefi, Baghayei, 2013) Therefore, if tests do not measure what intending to measure but something more resulting an invalid source for interpretation, the tests are bias. They surely decline all educational and social institution reputations. Their graduated students could not do a working program of their admitted job because they do not have the required ability and knowledge. Meanwhile, qualified individuals may be rejected and expelled from their deserved positions and rights.

Rater bias will affect a rater's judgment and can cause to systematic and continues errors in scoring, obscuring the accuracy of scores assigned. For example, a rater may systematically assign higher scores based on familiarity of the classroom setting or characteristics of the teacher, which can miss the interpretation and credibility of the performance category indicated by the scoring rubric. (Park, Chen, Holtzman, 2014) Thus, to improve the consistency and minimize rating errors, Janssen suggested that raters need to (1) be familiar with the measures they are using, (2) understand the sequence of operation, and (3) be trained on how they should interpret the scoring rubric (Janssen, 2015). Furthermore, other studies found that the background of the raters (trained or untrained) did not affect their reliability because completely change the rater's paradigm is not easy. (Brown, 1995; Eckes, 2008; Carey, Mannell, Dunn, 2011)

In rating writing performance, a rater can create a large variability in scoring. For example, in a classic study by Diederich, French, and Carlton (1961) in which three hundred essays were judged by fifty-three raters on five scoring criteria – ideas, form, flavor, mechanics, wording. Ideas means relevance, clarity, quantity, development, persuasiveness. Form is organization and analysis. Flavor reveals style, interest, sincerity. Mechanics is specific errors in grammar, punctuation, etc. and wording is choice and arrangement of words. Diederich et al. found that 94 percent of the essays received at least seven different

scores in rater severity. These differences become the factor that leads to differences in scores assigned, where some raters are more stringent or lenient than other raters – raters bias. (Diederich, French, Carlton, 1961) It is a systematic pattern of rater behavior that manifests itself in unusually severe or lenient ratings associated with a particular aspect of the assessment situation. (Eckes, 2012) Raters can score high or low degrees of severity when scoring writing performance of a certain group of students using a particular scoring criterion. Bachman and McManara stated that it is called bias analysis when raters show evidence of exercising this kind of differential severity (stringency or leniency) which exhibit differential rater functioning. (Bachman, 2014; McNamara, 1996) Other study revealed differences in raters to scoring precision in term of how well raters are able to discriminate differences between categories of the scoring (DeCarlo, 2005). As a result, when raters have lower scoring precision, they cannot discriminate differences between a high or a low score, and this can distract the real meaning of their scores.

Thomas Eckes (2005)in his study about *Rater types in writing performance assessments: A classifcation approach to rater*, differed raters significantly in their views based on the general importance of nine routinely scoring criteria in a TestDaF scoring rubric – fluency, train of thought, scructure, completeness, description, argumentation, syntax, vocabulary, correctness. *Fluency* is the degree to which the text can be read fluently. *Train of thought* is the degree to which the trained thought can be followed. *Structure* explains the degree to which the text is structured. *Completeness* reveals the degree to which all of the points specified in the task description are dealt with. *Description* means the degree to which the information contained in the prompt, such as a table or diagram, is summarized. A*rgumentation* is the degree to which points of view/personal considerations are recognizable. *Syntax* is the degree to which the text exhibits a range of cohesive elements and syntactic structures. He found that (a) raters differed markedly in the severity with their rating examinees, (b) raters were fairly consistent in their overall ratings, and (c) raters were substantially less consistent in relation to scoring criteria than in relation to examinees. In accordance with these findings, he revealed that rater perception and background of the scoring criteria may inclined scoring tendency and bias. (Eckes, 2005)

At the research of *Assessing EFL University Students' Writing: A Study of Score Reliability*, Quintero et al (2017) reports on the raters' views on writing assessment and their use of analytical scoring rubrics. They found that great variability was found between scoring criteria and raters differed in their levels of leniency and severity which means there is

scoring bias. They uncovered that having relatively similar background, perception and using the same rubric are not enough to ensure rater reliability. Raters' perceptions about the rubric determined their scoring tendency. They obtained their research data from five writing samples, adapted analitical scoring rubric and a rater background questionnaire. (Quintero, Guzmán, Guzmán, 2017)

Therefore, in this reseach I challanged the argument of Eckes (2005) that rater perception and background of the scoring criteria may inclined scoring tendency and bias, and Quentero et al (2007) about that having relatively similar background, perception and using the same rubric are not enough to ensure rater reliability. I propose that the raters' background, perception of the test construct, perception of the scoring rubric and using the same rubric would be sufficient to confirm the rater's reliability, if I combine all the components of raters' background, perception of the scoring rubric, using the same scoring rubric and statistical analysis of numerical data. (Sokolov, 2014) I assume that the result can give relevant information on the education setting in which the teacher or raters operate. (Koretz, 2008)

Therefore in this study, the combination of raters' background, perception of the scoring rubric, using the same scoring rubric and statistical analysis of numerical data were applied together. I used an analytical scoring rubric (Jacobs, Zinkgraf, Wormuth, Hartfiel, Hughey, 1981) in order to be used simultaneusly during rating task. I also adepted a survey for raters' background and questionairre of raters' perception of the scoring rubric (Park, Chen, Holtzman, 2014) to see the effect of the raters's background and perception in their rating task. Finally I applied Two-Way Anova, One-Way Anova and Hoyt's Anova to measure the raters' bias, background and perception in awarding score of writing performance. I invited seven voluntary raters having master degree in ELT and Linguistics to score three opinion essays written by three undergraduate students from English Department using the analitical writing rubric. (Jacobs et al., 1981) Thus, for the research questions of my study are as followed:

1. Is there any bias performed by 7 raters, particularly in content, organisation, vocabulary, language use and mechanis?
2. How does the raters' background affect their scoring, in term of having experience of writing lecturer and length of lecturing writing?
3. How does the raters' perception on their understanding rubric affect their scoring?

METHOD

I invited all 15 students of batch 2016's doctoral degree students of Teaching English in Universitas Negeri Malang, Indonesia and only seven of them were willing sincerely to be the subject of this research. They had master degree in Linguistics or English Language Teaching and were at their second semester when this study was conducted. The instruments of this reseach are an analitical writing rubric (ESL composition profile), three opinion essays, a questionairre of the raters' background (apendix 2) and a questionairre of perseption (apendix 3). Both questionairres are adapted with the questionnaires of rater background variables and rater survey variables used by Park et al. at their study about evaluation efforts to minimize rater bias in scoring classrom observations. (Park, Chen, Holtzman, 2014) The analitical writing rubric which was to measure opinion essays was adopted from the ESL compositon profile at the book "*Testing ESL Composition: a practical approach"* by Jacobs et al. (Jacobs, Zinkgraf, Wormuth, Hartfiel, Hughey, 1981) This rubric has been used by many previous studies. (Klein, 1987; Baak, 1997; Cahyono, 2000; Bacha, 2001; Nakanishi, 2005; Haswell, 2007; Ghanbari, Barati, Moinzadeh, 2012; Crusan, 2013; Jonathan Trace, 2017)

The opinion essays were written by three undergraduate students of English Department of Universitas Negeri Malang which valuntarily agreed to be the subject of my study. The raters scored three opinion essay based on the rubric and answered the questionairres of rater background and perception. There was no training for raters to score the essay because they were all master degree in Linguistics or English Language Teaching. To measure the bias on their scoring essay, I used Two-Way Anova which is variance stastitical assessment of reliability, utilizing the analysis of variance formulae. (Stuart, Halmilton, 2007) Next I also computed a numerical indicator of the reliability of the test and is called Hoyt's reliability (Clark, 1999). To process both Anovas, I used SPSS 22. I found that the scoring criteria of Content, Organization and Vocabulay rated by the raters were respectively were not reliable (0.195, 0.511, 0.545 $<$ 0,600). It means that the seven raters' judgement over the Content, Organization and Vocabulary was bias because the raters had relatively differential severity/liniency toward the mentioned scoring criteria. However for Language Use and Mechanics, the raters shared relatively the same thought or had high severity in both Language Use and Mechanics in scoring the essay, in term of not bias (0.660, 0.809 > 06.00).

Next I analyzed the seven raters' background that influence the opinion essays' scoring with One-Way Anova α 0.050 as the signifincant criteria. If the significant value is above 0.050, the raters's background (experience of teaching writing and length of teaching writing experience) have no effect on the raters' scoring rubric items (Content, Organisation, Vocabulary, Language Use and Mechanics), in term not bias. However, if the value is equal or below 0.050, the raters' background effect the scoring which is bias. I computed firstly the raters' background of having experience of writing lecturer and only the Mechanics' criteria of scoring was affected with the significant value 0.026 which is bias (0.026 $<$ 0.050). Another categories – Content, Organisation, Vocabulary and Language Use were not affected by the mentioned background because the significant values were above 0.050 (0.828, 0.913, 0.126, 0.093 $>$ 0.050). Thus, based on the background of having writing lecturer's experience, the seven raters's scoring only affected the Mechanics'criteria which was called bias because they had the different scoring result. Thus whether the raters have experiance of lecturing writing or not, their criteria scoring of Content, Organization, Vocabulary, Language Use were not affected in scoring here. So their experience in lecturing writing determined their scoring in Content, Organization, Vocabulary, and Language Use. But the seven raters had different thought in scoring Mechanics, no matter having lecturing experience of not. Brown (1995), Eckes (2008) and Carey et al. (2011) also found that the length may help to improve the raters' scoring quality, however to completely change the rater's paradigm is not easy and may take some times. (Brown, 1995; Eckes, 2008; Carey, Mannell, Dunn, 2011)

Then, I calculated the effect of the length of teaching writing experience to their scoring and found out that the length of lecturing writing experience did not affect at all to the scoring criteria of Content, Organization, Vocabulary, Language Use and Mechanics, in term of no bias in scoring at all (0.705, 0.663, 0.171, 0.206, 0.090 $\geq$ 0.050). Thus, no matter how long the raters had lecturing writing experience, their rating task using the scoring criteria was not affected. In other word, their length of lecturing esperience did not qualify thier rating task. As a result, based on the raters' background of having writing lecturers' experience, their scoring was bias in rating the Mechanics' criteria while for another criterias were not bias. In judging Mechanics criteria, here they had different scoring while for Content, Organization, Vocabulary, and Language Use they had relatively the same judgement in rating the three essays. Finally, in accordance with the background of the length of lecturing writing's experiance, their rating task using the five scoring criteria was not

affected. Indeed, in judging the quality of the essays based on the criteria of Content, Organization, Vocabulary, Language Use and Mechanics, the raters were not influanced by the lenght of lecturing writing's experiance.

Next I adopted the questionairre of Part et al (2014) to measure the seven raters' perception toward the used rubric. (Park, Chen, Holtzman, 2014). See appendix 3. This questionairre had been used by Park et al in their reseach about evaluating effort to minimize rater bias in scoring classromm observation. The first question of the questionairre was about the raters' familiarity toward the instrument of writing rubric prior to this research. The raters were required to select "*yes*" or "*no*". Meanwhile the question number two until eight of the questionairre was about the raters' perception and opinion about the scoring rubric that required raters to indicate the level of their agreement on a 5-point scale ranging from strongly disagree to strongly disagree.

I also used the rubric created by Jacobs et al because it had been used extensively in writing research. Finally I asked the raters to fill the questionnaire Based on their perception of acknowleging the rubric of ESL writing compositon (Jacobs, Zinkgraf, Wormuth, Hartfiel, Hughey, 1981). All seven raters admitted familiar with the instrument of writing rubric prior to this reseach. They agree that the rubric help them to discriminate among the different score level. They considered that the rangefinders in the rubric were usefull tools for understanding how to asign score and the writing rubric measured some of the essential elements for effectively teaching and learning writing. They deemed the valid and fair rubric could be used as a professional development tool to support or improve teaching and learning writing, and finally they were confident in their ability to score using the provided rubric.

RESULTS AND DISCUSSION

Measuring Raters' Bias in Scoring Writing Performance

After each of the seven raters measuring the three opinion essay based on the analitical writing rubric items (Content, Organisation, Vocabulary, Language Use and Mechanics), see appendix 1, I examined whether there was a bias on scoring the three opinion essays by seven raters. I computed the result (appendix 1) with Two-Way Anova to seek the mean square of the raters and students' essay as the independent variable, based on each dependent variable (Content, Organization, Vocabulary, Language Use and Mechanics).

The mean square value of each dependent variable then was computed with Hoyt's Anova of reliability test, the analysis of variance formulae in order to find out the reliabilty of rating task of the seven raters in measuring the three essays. The computation result which is not reliable, is named bias. I defined 0.600 as coefficient realiability standard. The classification of coefficient reliability is as followed: 0.800 - 1.000 = excellent , 0.600 – 0.799 = good, 0.400 – 0.599 = adequate, below 0.399 = may have limited applicability. (H-R Guide, 2015) Thus, if the result of Anova Hoyt's reliability test is equal or above 0.600, the scoring judgement amongst the seven raters is reliable, in term of not bias. However if it is below 0.600, the result is not reliable, which is meant bias. Below is the computation of Two-Way Anova of all Criteria and Hoyt'Anova of realibility to measure the scoring reliability amongst the seven raters according to each dependent variable: Content, Organization, Vocabulary, Language Use, Mechanics.

Table 1. Reliability test result of Hoyt's Anova for 5 independet variable judged by 7 raters

No.	Rubric Criteria / Dependent Variables	Result of Hoyt's Anova test	Standard of Reliability interpretation	Quality of reliability interpretation	Rater's Scoring Interpretation
1.	Content	0.195	≥ 0.600	not reliable	Bias
2.	Organization	0.511	≥ 0.600	not reliable	Bias
3.	Vocabulary	0.545	≥ 0.600	not reliable	Bias
4.	Language Use	0.660	≥ 0.600	reliable	Not bias
5.	Mechanics	0.809	≥ 0.600	reliable	Not bias
	Total	0.514	≥ 0.600	not reliable	Bias

Based on the Hoyt's Anova test, the seven raters scored Language Use and Mechanics well with no bias. (0.660 and 0.809 ≥ 0.600) It means the seven raters had relatively the same point of view in measuring the three essay based on the Language Use and Mechanics items, although only three of them having experience of lecturing writing. Thus in measuring effective complex construction, agreement, tenses, number, word order/function, articles, pronouns and prepositions for Language Use and assessing language convention of spelling, punctuation, capitalization and paragraphing for Mechanics, (Jacobs, Zinkgraf, Wormuth, Hartfiel, Hughey, 1981) they had relatively valid rating consistency, but for the other criteria their rating was not reliable which means bias. The study of *Examining Rater Effects in TestDaf Writing and Speaking Performance Assessment: A Many-Facet Rasch Analysis*, Thomas Eckes also found that raters had different strongly in severity in assessing the students' writing and speaking performance and are substantially less consistent in relation to

criteria of writing rubric. (Eckes, 2005) Schaefer's (2008) study also yielded valuable insight into specific patterns of bias shared by subgroups of raters. (Schaefer, 2008) Meanwhile in this study, I found that the scoring criteria of Content, Organization and Vocabulay awared by the raters were respectively 0.195, 0.511, 0.545 < 0.600. It means that the seven raters' judgement over the Content, Organization and Vocabulary were not reliable, in term of bias because the raters had relatively differential severity/liniency toward the mentioned coring criteria. However for Language Use and Mechanics, the raters shared relatively the same thought in scoring the essay or not bias (0.660, 0.809 > 06.00). Since they have high severity in both Language Use and Mechanics, I consider them as novice writer.

Effect of Raters' background in their scoring

Next I scrutinized the raters' background effect to their rating task by using One Way Anova of which the computation was processed with SPSS 22. Previously, the seven raters answered the rater's background variable questionairre (see apendix 2) which is consisted of gender, race/ethnicity, having experience of writing lecturer, length of being writing lecturer and highest degree of the raters. There were four males and three females coming from some islands in Indonesia like Riau, Sumatra, Java, Madura and Kalimantan. Out of the seven raters, one male was from Libya. Four raters had no eperience of lecturing writing and three of them had ever been teaching ranging from three to ten years. These seven raters had the same master degree in Linguistics and English Language Teaching.

In this reseach, I only used the variables of having experience of writing lecturer and length of writing lecture's experience, to be paralled with the scoring precision using the analitical writing rubric in order to know whether the background affecting the result of the rating task. In other words I wanted to know how well the seven raters were able to discriminate differences between categories of the scoring (DeCarlo, 2005). If they had lower scoring exactness, they could not distinguish the differences between a high or a low score, and this may divert the real essense of measuring writing performance. (DeCarlo, 2005) Congdon et al (2000) stated that rater's understanding of the writing's rubric profile and measurement may reduce bias and variance in scoring system which improve consistency and minimize rating errors. (Congdon, McQueen, 2000) Meanwhile I assumed the gender and race/ethnicity gave no signifince effect in their scoring effect. For the evidence of gender bias, Thomas Eckes found that the calibration values for the gender facet were either very

small (and not significantly different), indicating gender bias favoring men, or very large (and significantly different), indicating gender bias favoring women. (Eckes, 2005)

Fistly, I analyzed the variable of having experience of writing lecturer with the analitical rubric categories using with One-Way Anova $\propto$ 0.050 as the significant criterion or p-value. It means that if the result of the rubric items is equal to or above 0.050, the raters' background of the writing lecturer's experience influences the raters' scoring rubric items (Content, Organisation, Vocabulary, Language Use and Mechanics). However if the result is below the criterion, the background does not influance the raters' scoring judgement. Below is the One-Way Anova computation of the raters' background of having experience of writing lecturer which affects the raters' scoring.

Table 2. One Way Anova test for Independent variable: Having Experience of Writing Lecturer

Independent Variable	Dependent Variable	Sum of Squares	Degree of freedom	Mean Square	F value	Signifincant Value
Having Experience of writing lecturer	Content	.397	1	.397	.048	.828
	Organization	.099	1	.099	.012	.913
	Vocabulary	13.349	1	13.349	2.556	.126
	Language Use	20.004	1	20.004	3.138	.093
	Mechanics	.893	1	.893	5.816	.026
	Total	66.036	1	66.036	.797	.383

At table 2, I used One-Way Anova to compute the effect of having background of experience of writing lecturer as the independent variable to their scoring criteria of Content, Organization, Vocabulaty, Language Use and Mechanics as dependent variable and processed the computation with SPSS 22. I found out that the significant value of dependent variables were 0.828, 0913, 0.126, 0.093, 0.026 respectively. Next I put my finding in table 3.

Table 3. Effect of Writing Lecturer's Experience to Raters' scoring by One-Way Anova test

No	Dependent Variables	Significance Value	Significance Criteria / p-value	Quality of affect interpretation	Raters' Scoring Interpretation
1	Content	0.828	$\geq$ 0.050	Not affected	Not bias
2	Organization	0.913	$\geq$ 0.050	Not affected	Not bias
3	Vocabulary	0.126	$\geq$ 0.050	Not affected	Not bias
4	Language Use	0.093	$\geq$ 0.050	Not affected	Not bias
5	Mechanics	0.026	$\geq$ 0.050	Affected	bias
	Total	0.383	$\geq$ 0.050	Not affected	Not bias

.

Table 3 is the raters' background variable of Mechanics having experience of writing lecturer which affecting the variability of raters' scoring. Only the rublic item of was affected or bias. (0.026 < 0.050) The seven raters only had relatively different scoring judgement in Mechanics. Meanwhile, the Content, Organization, Vacabulary and Language Use were not affected by the independent variable or respectively 0.705, 0.663, 0.171, 0.206 > 0.050 because there were a relatively the same variability in awarding the scores based on these criterions. In line with this background, the seven raters judgement was not affected, thus, in awarding the scores to the three the essays based on the Content, Organization, Vocabulary and Language Use was not bias. It means they shared reliable scoring according to these criterions. In general, the raters' background of obtaining expeerience of writing lecturer was not bias 0.383 > 0.050 because they used the same scoring rubric to ensure the raters' reliability. Benefits of using scoring rubric in writing performance assessments would increase consistency of scoring since it possibly facilitated valid judgment of complex competencies and criteria explicitly. (Jonsson, Svingby, 2007)

At table 4, by using One-Way Anova with SPSS 22, I computed the effect of background of how long the raters lecturing writing as the independent

Table 4. One Way Anova test for Independent variable: **Length of Lecturing Writing**

Independent Variable	Dependent Variable	Sum of Squares	Degree of freedom	Mean Square	F value	Significant Value
Length of Lecturing Writing	Content	5.952	2	2.976	.356	.705
	Organization	6.821	2	3.411	.421	.663
	Vocabulary	20.071	2	10.036	1.953	.171
	Language Use	22.726	2	11.363	1.727	.206
	Mechanics	.893	2	.446	2.755	.090
	Total	150.536	2	75.268	.909	.421

variable to their scoring judgement based on criteria of Content, Organization, Vocabulary, Language Use, Mechanics as dependent variables. The computation yielded the significant value of dependent variables were 0.705, 0.663, 0.171, 0.206, 0.090. Then I placed these sifnificant value in table 11 in order to be interpreted.

Table 5. Effect of Length of Lecturing writing to Raters' scoring by One-Way Anova test

No	Dependent Variables	Significance Value	Significance Criteria / p-value	Quality of affect interpretation	Raters' Scoring Interpretation

1	Content	0.705	≥ 0.050	Not affected	Not bias
2	Organization	0.663	≥ 0.050	Not affected	Not bias
3	Vocabulary	0.171	≥ 0.050	Not affected	Not bias
4	Language Use	0.206	≥ 0.050	Not affected	Not bias
5	Mechanics	0.090	≥ 0.050	Not affected	Not bias
	Total	0.421	≥ 0.050	Not affected	Not bias

Table 5. explained the background of length of lecturing writing did not affect the raters' judgement in scoring the essay > 0.050. Thus in general (0.421 > 0.050) no matter how long they teach writing or even not having experience of lecturing writing, they rated the essays reliably by using the same analitical writing rubric. It also indicated that using the same analitical rubric of ESL writing composition may improve the reliability of their rating although four raters in this study were not well trained on how to design and employ them effectively. (Rezaei, Lovorn, 2010) This can be caused by their similar perception to the rubric used in this study and it would be explain in the next describtion.

Effect of Raters' perception on understanding the rubric used in their scoring

Below is the perception of the seven raters towards the analytical rubric presented in descritive statistics. They had answered the questionairre of the rubric perception and here are the result of their answers.

Table 6 A. Raters' Familiarity Perception on the scoring rubric

No.	Questions	Answer	
		Yes	No
1	Were you familiar with the instrument of writing rubric prior to this reseach?	7 (100%)	-

The first question was "Were you familiar with the instrument of writing rubric prior to this research?" The seven raters answered yes. This is because all of them are having master degree in Linguistics and Teaching English Language. Although four of the researchers have no experience in lecturing writing, such this rubric become their tacit knowledge.

Table 6 B. Raters' Familiarity Perception on the scoring rubric

No.	Question	Answer				
		SD*	D*	N*	A*	SA*
2	I was confident in my ability to score using the provided rubric.	-	1	1	4	1
3	The rubrics were clear and helped me to discriminate among the different score levels	-	-	-	6	1
4	The benchmarks and rangefinders in the writing rubric were	-	-	-	5	2

	useful tools for understanding how to assign scores.					
5	The writing rubric measures some of the essential elements for effectively teaching Writing.	-	-	-	6	1
6	The writing rubric measures some of the essential elements for effectively learning Writing.	-	-	-	5	2
7	The writing rubric could be used as a professional development tool to support or improve teaching and learning writing.	-	-	1	6	-
8	The writing rubric is a fair and valid teaching observation tool.	-	-	2	5	-

- SD = Strongly disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly agree

The perception of the raters for the question number two "*I was confident in my ability to score using the provided rubric*" mostly agree as the four raters answered "*agree*" while the other three raters gave answer "*disagree*", "neutral", "*strongly agree*". The next question about the clarity of the rubric of analytical writing to help the raters discriminating among the different score level, the six raters replied "*agree*" and one rater answered "*strongly disagree*". They thought that the analitical rubric really assisted them to focus on each of various assigned aspects of the writing sample, so that they all evaluated the same features of a student's performance. This finding is in line with the study by Nakamura (2004) and Jonsson et al. (2007) that one of the advantages of analitical rubric, the raters can avoid the risk of idiosyncratic in their rating task. Meanwhile, for the question number four "*The benchmarks and rangefinders in the writing rubric were useful tools for understanding how to assign scores*" five raters answer "*agree*" and two raters answer "*strongly agree*". The benchmark and rangefinder can increase inter-rater reliability, besides avoiding the risk of idiosyncratic when they award the score to the test takers' responses. The analytical or specific topic rubric performance will enhance the scoring reliability of writing performance assessments. (Nakamura, 2004; Jonsson, Svingby, 2007)

The perception questionairre number five about "*The writing rubric measures some of the essential elements for effectively teaching Writing*", six rater answered "*agree*" and one chosed "*strongly agree*". In line with the writing rubric measuring some of the essential elements for effectively learning Writing for question number six, five raters selected "*agree*" and two raters opted "*strongly agree*". It means the seven raters had known that the fuction of the rubric is both for teaching and learning, particularly the analytical rubric used in this research. They realize the rubric can assist them to teach and evaluate the students' work well. Some reseachers Kacy Lundstrom, Anne R. Diekema, Heather Leary, Sheri Haderlie, Wendy Holliday (2015) in their study about "*Teaching and Learning Information Synthesis*" found that the rubric benefits the students to synthesize their lesson. Although the level of

synthesis low in overall, they could identify different levels of information integration. These reseachers discovered that the rubric is effective ways to measure and teach synthesis which were essential in helping students become information literate. (Lundstrom, Diekema, Leary, Haderlie, Holliday, 2015)

The question number seven at the questionairre, the six raters answered “agree” and one rater chosed “neutral” about the statement “The writing rubric could be used as a professional development tool to support or improve teaching and learning writing”. Last but not least, the last question “The writing rubric is a fair and valid teaching observation tool”, five raters stated “agree” while two raters prefered “normal”. The seven raters considered that the rubric is good for professional development in order to accelerate teaching and learning writing. For experienced raters, rubric define critical dimensions of teaching as the basis of the evaluation for salary increment, and other forms of teacher recognition, such as the selection of mentor or lead teachers. (Hammond, 2010) In addition to that, they also thought that the applied rubric was fair and valid teaching observation tool which describe levels of performance during self-assessment. (NC Departement of Public Instruction, 2015)

Discussion

Since human scoring is definetely subjective and inclining to bias, (Schaefer, E. , 2008), among the five rubric criteria (Jacobs et al, 1981) being analizing with Two-Way Anova and Hoyt reliability test, only Language Use and Mechanics were not bias (0.660 and 0.809 $\geq$ 0,600). The raters had relatively the same point of view in awarding the scores for the three essay based on the grammar and convention criteria. Thus, the raters’ perceptions of grammar had a predominant influence on awarding test scores. (McNamara, 1996) However, the other criteria – Content, Organization and Vocabulary, their rating was bias which meant each rater had differential severity in rating the three essay (0.195, 0.511, 0.545 $\geq$ 0,600). Eckes (2005) and Schaefer (2008) also found that raters had different strongly in severity in assessing the students’ writing performance and were substantially less consistent in relation to criteria of writing rubric. (Eckes, 2005; Schaefer, 2008)

In this study, the raters had master degree in Linguistics and English language teaching and they focused and shared relatively the same severity in scoring criteria of Language Use and Mechanics refering to bottom-up feature. Highly proficient raters tended to use a top-down approach to essay scoring, focusing on performance features that are more general. Less proficient raters tended to use a bottom-up approach to essay

scoring, focusing on performance features that are more specific. Less proficient raters tended to use a bottom-up approach to essay scoring, focusing on performance features that are more specific. Highly proficient raters tended to use a top-down approach to essay scoring, focusing on performance features that are more general. (Eckes, 2012) Threfore, I consider the seven raters in my study less proficient raters because they tended to use a bottom-up approach to essay scoring, focusing on performance features that are more specific.

In this reseach, I only used the variables of having experience of writing lecturer and length of writing lecture's experience, to be paralled with the scoring precision using the analitical writing rubric. I explored how their background affecting the result of the rating task, in term of how well they discriminated differences between categories of the scoring (DeCarlo, 2005). If they had lower scoring exactness, they cannot distinguish the differences between a high or a low score, and this may divert the real essense of measuring writing performance. (DeCarlo, 2005) Meanwhile I assumed the gender and race/ethnicity gave no signifince effect in their scoring effect. For the evidence of gender bias, Thomas Eckes confirmed that the calibration values for the gender facet were either very small (and not significantly different), indicating gender bias favoring men, or very large (and significantly different), indicating gender bias favoring women. (Eckes, 2005)

By using with One-Way Anova and Hoyt's reliability test, I measured the raters' background variable of having experience of writing lecturer which affecting the variability of raters' scoring. Only the rublic item of Mechanics was affected or bias ($0.026 < 0.050$), meaning that they did not share reliable in Mechanics. But in general, the background having expeerience of writing lecturer did not effect their rating task in scoring the criteria of Content, Organization, Vocabulary and Language Use. This could be that they used the same scoring rubric to ensure the raters' reliability. Benefits of using scoring rubric in writing performance assessments would increase consistency of scoring since it possibly facilitated valid judgment of complex competencies and criteria explicitly. (Jonsson, Svingby, 2007)

Meanwhile, in this study, the background of length of lecturing writing could not effect the rating judgement. Thus in general ($0.421 > 0.050$) no matter how long they teach writing or even not having experience of lecturing writing, they rated the essays reliably by using the same analitical writing rubric. It also indicated that using the same analitical rubric of ESL writing composition may improve the reliability of their rating although four raters in this study were not well trained on how to design and employ them effectively. (Rezaei,

Lovorn, 2010) Thus, although four raters did not have experience in teaching writing, they were benefited by the scoring rubrics of writing performance assessments. The rubric could increase their consistency of scoring and the possibility to facilitate valid judgment of complex competencies. In this regard, Jonsson and Svingby (2007) found out that: (1) the reliable scoring of performance assessments can be enhanced, especially if the rubrics are analytic, topics specific,and complemented with exemplars and/or rater training; (2) rubrics do not facilitate valid judgment of performance assessments per se. However, valid assessment could be facilitated by using a more comprehensive framework of validity when validating the rubric. (Jonsson, Svingby, 2007) The main reason for this emimance is fact that the explicit expectations and criteria of the rubric will definately facilitates feedback and self-assessment, as well as professional development.

Additionally, the seven raters were familiar with the instrument of writing rubric prior to this study because they are having master degree in Linguistics and Teaching English Language. Although four of the researchers had no experience in lecturing writing, such this rubric become their tacit knowledge. Many disciplines also interest in rubrics since the content, focus, type of rubrics used, as well as the actors involved, also can be varied considerably. Therefoe the rubric userse range from K-12, college, and university to active professionals is represented. (Jonsson, Svingby, 2007)

Finally, all raters in this study were somehow confident in their ability to score using the provided rubric. About the clarity of the analytical writing rubric, they also agreed that it helped them to discriminate the different score levels. They also thought that the analitical rubric really assisted them to focus on each of various assigned aspects of the writing sample, so that they evaluated the same features of test takers' writing performance. This regard is in line with some researchers' studies like Nakamura (2004) and Jonsson et al. (2007) that one of the advantages of analitical rubric, the raters can avoid the risk of idiosyncratic in their rating task. The seven raters in this study also agreed that "*the benchmarks and rangefinders in the writing rubric were useful tools for understanding how to assign scores*" five raters answer "*agree*" and two raters answer "*strongly agree*". The benchmark and rangefinder can increase inter-rater reliability, besides avoiding the risk of idiosyncratic when they award the score to the test takers' responses. The analytical or specific topic rubric performance will enhance the scoring reliability of writing performance assessments. (Nakamura, 2004; Jonsson, Svingby, 2007) The seven raters also corresponded that the writing rubric measures some of the essential elements for effectively teaching

Writing. The seven raters, in fact, knew that the fuction of the rubric is both for teaching and learning, particularly the analytical rubric used in this research. They realize the rubric can assist them to teach and evaluate the students' work well. Some reseachers Kacy Lundstrom, Anne R. Diekema, Heather Leary, Sheri Haderlie, Wendy Holliday (2015) in their study about "*Teaching and Learning Information Synthesis*" found that the rubric benefits the students to synthesize their lesson. Although the level of synthesis is low in overall, they could identify different levels of information integration. These reseachers discovered that the rubric is effective ways to measure and teach synthesis which were essential in helping students become information literate. (Lundstrom, Diekema, Leary, Haderlie, Holliday, 2015)

Finally they approved that the writing rubric could be used as a professional development tool to support or improve teaching and learning writing. The applied rubric was considered good for professional development and can accelerate teaching and learning writing. For experienced raters, rubric define critical dimensions of teaching as the basis of the evaluation for salary increment, and other forms of teacher recognition, such as the selection of mentor or lead teachers. (Hammond, 2010) In addition to that, the raters also thought that the applied rubric was fair and valid teaching observation tool which describe levels of performance during self-assessment. (NC Departement of Public Instruction, 2015)

CONCLUSION

Human raters comitts to bias and have strongly difference in severity to assess the students' writing performance. They can be less consistent in relation to criteria of writing rubric. Bias happens due to interaction between raters and the criteria. Therefore, some raters can score more consistently or harshly on a criterion referring to grammar, fluency, vocabulary, language use and mechanics whereas others scored more leniently on these criteria. In this study, the scoring criteria of Content, Organization and Vocabulay rated by the raters were respectively were not reliable because they have different scoring. It means they had relatively differential severity/liniency towards the top-down approach-content, organization and vocabulary are the top.

On the other hand they have similar scoring toward language use and mechanics which are the bottom-up approach to score essay. They consider less proficient raters raters as they focus on performance features that are more specific. They seemed to be more likely to resort to the gramatical and convention and some other aspects not captured in the scoring rubric. Meanwhile proficient raters tended to use a top-down approach to essay scoring,

focusing on performance features that are more general. In general the novice raters had the same criteria as the expert raters had, but their attention were devoted to discovering the criteria, in term of not able to engage with the texts.

The raters' background of having experience of writing lecturer could affect the raters' scoring of Mechanics which means that they shared differential reliablility in Mechanics - bias. But the background having experience of writing lecturer did not effect their rating task in scoring the criteria of Content, Organization, Vocabulary and Language Use. This might be that they used the same scoring rubric which can increase the raters' reliability and facilitate valid judgment of complex competencies and criteria explicitly.

Additionally, the background of the length of lecturing writing could not effect the raters judgement to award the scores – not bias. Thus, no matter how long they ever teach writing or not, they rated the essays reliably by using the same analitical writing rubric. Therefore, using the same analitical rubric of ESL writing composition may improve the reliability of their rating. However, rubrics do not facilitate valid judgment of performance assessments, but, valid assessment could be facilitated by using a more comprehensive framework of validity when validating the rubric. The main reason for this emimance is the fact that the explicit expectations and criteria of the rubric will definately facilitates feedback and self-assessment, as well as professional development.

Since the seven raters were familiar with the instrument of writing rubric prior to this study because they are having master degree in Linguistics and Teaching English Language. Although four of the seven raters had no experience in lecturing writing, such this rubric become their tacit knowledge. As a result, they were somehow confident in their ability to score using the provided rubric. They have positive point of view about the analitical rubric components, including the clarity of discrimination, writing focus and feature to avoid idiosyncratic in their rating task and rangefinders as useful tools for understanding how to assign scores. Finally they approved that the writing rubric could be used as a professional development tool to support or improve teaching and learning writing. For experienced raters, rubric define critical dimensions of teaching as the basis of the evaluation for salary increment, and other forms of teacher recognition, such as the selection of mentor or leading teachers.

BIOPROFILE

Endah Yulia Rahayu is an ELT specialist who is keen on English language testing and writing assessment. She is on a study leave at Universitas Negeri Malang and is projected to finish in 2019. She works for English Department of Teacher Training Faculty Universitas PGRI Adi buana Surabaya, East Java, Indonesia. Recently she has been conducting her study intensively in rater behavior, perspective and scoring in writing performance. She is also interested in ELT management and leadership.

REFERENCES

Amin, I. A.-R., Aly, M. A.-S., & Amin, M. M. (2011). A Correlation Study between EFL Strategic Listening and Listening Comprehension Skills among Secondary School Studnets. Benha, Egypt: Benha University.

Baak, E. (1997). Portfolio Development: An Introduction. *Forum, 35*(2), 38.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, 371-383.

Bachman, L. F. (2014). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Basir, A. (2014). Autistic Students' Learning Strategies in Writing English Texts and Their Impacts on The Teaching and Learning Process. Surakarta: Sebelas Maret University.

Bill & Melinda Gates Foundation. (2012). Gathering feed- back for teaching: Combining high-quality observations with student surveys and achievement gains. *Measures of Effective Teaching (MET).* Seattle, WA: Author.

Bozorgian, H., & Pillay, H. (2013). Enhancing Foreign Language Learning through Listening Strategies Delivered in L1: An Experimental Study. *International Journal of Instruction, 6*(1), 105-122.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specifc language performance test. *Language Testing, 12*, 1-15.

Brown, H. D. (2006). *Teaching by Principles: An Interactive Approach to Language Pedagogy.* New Jersey: Prentice Hall Regents.

Cabaysa, C. C., & Baetiong, L. R. (2010). Language Learning Strategies of Students at Different Levels of Speaking Proficiency. *Education Quarterly, 61*(8), 16-35.

Cahyono, B. Y. (2000, August). The Overall Proficiency in English Composition of Indonesian: University Students of EFL. *TEFLIN Journal, 11*(1), 78-87.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral profciency interviews? *Language Testing, 28*, 201–219.

Celce-Murcia, M. (2001). *Teaching English as a Second or Foreign Language.* Boston: Heinle & Heinle Publishers.

Chang, C. Y., Liu, S., & Lee, Y. (2007). A study of language learning strategies used by college EFL learners in Taiwan. *Language Learning, 3*, 235-262.

Clark, K. (1999, November). Test Realibility. *The Mathematics Teacher*, *92*(8), 719-723.

Cohen, D. (1998). *Strategies in Learning and using a Second Language.* London: Longman.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163–178.

Coskun, A. (2010). The Effect of Metacognitive Strategy Training on the Listening Performance of Beginner Students. *Novitas-ROYAL (Research on Youth and Language), 4*(1), 35-50.

Crusan, D. (2013, November 14). Designing Writing Assessment and Rubrics LARC/CALPER Testing & Assessment Webinar. Dayton, OH, USA.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53–76.

Diederich, P.D., French, J.W., Carlton, S.T. (1961). *Factors in Judgements of Writing Ability.* Princeton, New Jersey: Educational Testing Service.

Eckes, T. (2005). Examining Rater Effects in TestDaf Writing and Speaking Performance Assessment. *Language Assessment Quarterly, 2*(3), 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: A classifcation approach to rater. *Language Testing, 25*(2), 155-185.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9*, 270-292.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9*, 270–292.

Ellis, R. (1994). *The Study of Second Language Acquisition.* New York: Oxford University Press.

Farlex. (2007). Retrieved January 20, 2016, from The Free Dictionary: http://thefreedictionary.com

Gestanti, R. A. (2015). *Students'Learning Strategies and Their Accomplishment in Speaking English.* Surakarta: Sebelas Maret University.

Ghaderpanahi, L. (2012). Using Authentic Aural Materials to Develop Listening Comprehension in the EFL Classroom. *English Language Teaching, 5*(6), 146-153.

Ghanbari, B., Barati, H., Moinzadeh, A. (2012). Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny. *Language Testing in Asia, 2*(1), 83-100.

Gilakjani, A. P., & M. R. (2011). A study of Factors Affecting EFL Learners' English Listening Comprehension and The Strategies for Improvement. *Journal of Language Teaching and Research, 2*(5), 977-988.

Gilakjani, A. P., & Sabouri, N. B. (2016). Learners' Listening Comprehension Difficulties in English Language. *English Language Teaching, 9*(6), 123-133.

Hammond, L.D. (2010). *Evaluating Teacher Effectiveness How Teacher Performance Assessments Can Measure and Improve Teaching.* Washington, DC: Center for American Progress.

Harmer, J. (2001). *The Practice of English Language Teaching.* New York: Longman.

Harmer, J. (2007). *How to Teach English: New Edition.* London: Pearson Education Limited.

Haswell, R. H. (2007, January 15). Researching Teacher Evaluation of Second Language Writing via Prototype Theory. Corpus Christi, Texas, USA.

H-R Guide. (2015, May 12). *Chapter 3: Understanding Test Quality-Concepts of Reliability and Validity*. Retrieved 2017, from Human Resources: http://www.hr-guide.com/data/G362.htm

Huang, Y. F. (2009). The Relationship between College Students' Learning Strategies and Their English Speaking Proficiency. Ming Chuan, : Ming Chuan Univ Press.

Huda, N. (1998). Relationship between Speaking Proficiency, Reflectivity-impulsivity, and L2 Learning Strategies. *Learners and Language Learning. RELC Anthology series, 39*, 40-45. (W. Renandya, & G. M. Jacobs, Eds.) Singapore: SEAMEO Regional Language Centre.

Huy, L. H. (2015). An Investigation into Listening Strategies of EFL Students. *Asian Journal of Educational Research, 3*(4), 21-34.

Ivarsson, E., & Palm, M. (2013). Listening Strategies in the L2 Classroom. Malmö högskola.

Jacobs, H.L., Zinkgraf S.A., Wormuth D.R., Hartfiel V.F., Hughey J.B. (1981). *Testing ESL Composition: a practical approach.* Rowley, Massachusetts: Newbury House.

Janssen, F. J. (2015, January). *Research on rater bias in classroom observation*. Retrieved March 23, 2017, from http://janbri.nl/?page_id=103

Jonathan Trace, G. J. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22.

Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 130–144.

Jou, Y.-J. (2009). A Study of English Listening Strategies Applied by. Cheng Shiu: Cheng Shiu University.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Masurement, 1*, 1-73.

Kassem, H. M. (2015). The Relation between Listening Strategies Used by Egyptian EFL College Sophomores and Their Listening Comprehension and Self-Efficacy. *English Language Teaching Journal, 8*(2), 153-169.

Khamdani, A. K. (2014). *Learning Strategies Applied by Students of Nursing Academy in Listening.* Surakarta: Sebelas Maret University.

Klein, C. R. (1987). The value of assignment-specific writing scales for ESL composition. Ames, Iowa, USA.

Knoch, U., Fairbairn, J., Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in Language Testing and Assessment* , *5*(1), 90-106.

Knoch, U., Read, J., von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing, 12*, 26-43.

Koretz, D. (2008). *Measuring up: What educational testing really tell us.* Massachusetts/London, England: Harvard University Press.

Liang, T. (2009). Language Learning Strategies- The Theoretical Framework and Some Suggestions for Learner Training Practice. *English Language Teaching Journa*, *2*(4), 199-206.

Lundstrom, K., Diekema, A. R., Leary, H., Haderlie, S., Holliday. W. (2015). Teaching and learning information synthesis. *Communications in Information Literacy, 9*(1), 60-82.

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

McNamara, T., Roever, C. (2006). *Language Testing: The Social Dimension.* Oxford: Blackwell Publishing.

Meier, V. (n.d.). *Evaluating rater and rubric performance on writng placement exam.* University of Hawai‘i at Mānoa.

Milles, M. B., & Huberman, A. M. (1984). *Qualitative Data Analysis. A Sourcebook of New Methods.* California: SAGE Publication, Inc.

Nakamura, Y. (2004). A comparison of holistic and analytic scoring methods in the assessment of writing. *Proceedings of the 3rd Annual JALT Pan-SIG Conference* (pp. 45-52). Tokyo: 2004 Pan SIG.

Nakanishi, C. (2005). What Influences the Quality of Japanese College Students' Writing in English as a Foreign Language? *The Journal of Asia TEFL , 2*(1), 155-180.

Nation, I., & Newton, J. (2009). *Teaching ESL/EFL Listening and Speaking.* New York: Routledge.

NC Departement of Public Instruction. (2015). *North Carolina Teacher Evaluation Process.* Raleigh: Public School of North Carolina.

O'Malley, J. M. (1990). *Learning Strategies in Second Language Acquisition.* Cambridge: Cambridge University Press.

Oxford, R. L. (1990). *Language Learning Strategies: What Every Teacher Should Know.* Boston: Heinle & Heinle Publishers .

Oxford, R. L. (2003). *Language Learning Styles and Strategies: An Overview.* London: GALA.

Park, Y.S., Chen, J., Holtzman, S.L. (2014). Evaluating Efforts to Minimize Rater Bias in Scoring Classroom Observations. In T. K. Kane, *DESIGNING TEACHER EVALUATION SYSTEMS* (p. 384). San Francisco: Wiley.

Penulis, T. (2015). *Panduan Akademik 2015/2016.* Ponorogo: UMP Press.

Quintero, E.F.G., Guzmán, N.P.T, Guzmán, R.R. (2017). Assessing EFL University Students' Writing: A Study of Score Reliability. *Revista Electrónica de Investigación Educativa, 9*(2).

Razawi, N. A. (2011). Students' Diverse Learning Styles in Learning English as a Second Language. *International Journal of Bussiness and Social Science, 2*(19), 179-186.

Rezaei, A.R., Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.

Richards, J. C. (2002). *Methodology in Language Teaching. An Anthology of Current Practice.* New York: Cambridge University Press.

Riduwan. (2004). *Metode dan Teknik Menyusun Thesis.* Bandung: Alfabeta.

Rost, M. (1994). *Introducing Listening.* London: Penguin Group.

Saeidi, M., Yousefi, M., Baghayei, P. (2013). Rater Bias in Assessing Iranian EFL Learners' Writing Performance. *Iranian Journal of Applied Linguistics (IJAL), 16*(1), 145-175.

Schaefer, E. . (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493.

Shi, C. (2011). A Study of the Relationship between Cognitive Styles and. *Higher Education Studies, 1*(1), 20-26.

Shutler, J. (2002, August 15). One way ANOVA - Analysis of variance. Retrieved April 21, 2017

Sokolov, C. (2014). Self-evaluation of rater bias in written composition assessment. *Linguistica, 54*(1), 261-275.

Stuart, I., Halmilton. (2007). *Dictionary of Psychological Testing, Assessment and Treatment* (second ed.). London and Philadelphia: Jessica Kingsley .

Sudweeks, R. R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and Many-Facet Rasch Measurement in an analysis of college sophomore writing. *Assessing Writing*, 239-261.

Underwood, M. (1989). *Teaching Listening.* London: Longman.

Ur, P. (1996). *A Course in Language Teaching Practice and Theory.* Melbourne: Cambridge University Press.

Watthajarukiat, T. E. (2011). An Investigation of English Listening Strategies Used by Thai Undergraduate Students in Public Universities in the South Thailand. *Journal of Art, 15*(4), 1-17.

Weir, J. C. (1998). *Communicative Language Testing.* New Jersey: Prentice Hall Europe.

Wenden, A. &. (1987). *Learner Strategies in Language Learning.* New Jersey: Prentice Hall.

Wigglesworth, G. . (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

Zare, P. (2012). Language Learning Strategies among EFL/ESL Learners: A Review of Literature. *International Journal of Humanities and Social Science, 2*(5), 162-169.

Zhang, W.-S. (2007). Teach More Strategiesin EFL CollegeListening Classroom. *US-China Education Review, 4*(3), 71-76.

Appendices are upon request