UNIVERSITAS ISLAM MALANG

International Conference

ELTAR

English Language Teaching and Research

PROCEEDINGS

ELTAR 2018

International Conference
on Language, Education, and Culture

February 21 - 22, 2018
Ali bin Abi Thalib Building
Universitas Islam Malang

PROCEEDINGS

International Seminar on English Language Teaching and Research

Malang, February 21-22, 2018
Postgraduate Program
Universitas Islam Malang

PROCEEDINGS

INTERNATIONAL SEMINAR ON ENGLISH LANGUAGE TEACHING AND RESEARCH

Malang, February 21-22, 2018
Postgraduate Program
Universitas Islam Malang

Keynote Speakers:

Prof. Roderick J. Ellis, Ph.D.	School of Education, Curtin University, Australia
Dr. Anitha Devi Pillai	National Institute of Education Singapore
Prof. Dr. Gunadi H. Sulistyo, M.A.	Universitas Negeri Malang, Indonesia
Prof. Drs. H. Junaidi Mistar, M.Pd, Ph.D.	Universitas Islam Malang, Indonesia
Prof. Aslam Khan bin Samahs Khan, Ph.D.	Erican College, Malaysia

Welcome Address by the Rector of University of Islam Malang
Prof. Dr. H. Maskuri, M.Si

Distinguished Keynote Speakers
Distinguished Conference Presenters and Participants
Ladies and Gentlemen

Assalamu'alaikum War. Wab.
Good morning to you all.
First of all, let us bow our heads to express our gratitude to Allah SWT the Almighty, the Most Merciful and the Most Beneficial, for granting us His blessings and grace. Our deepest gratitude should also go to the Prophet Muhammad SAW who has led human kinds from darkness to lightness with the teachings of Islam.

At this very moment, on behalf of the big family of the University of Islam Malang, I would like to extend my warmest regards and greetings to the invited keynote speakers, presenters, and all participants of this first conference on English Language Teaching and Research (ELTAR), carried out by English Language Education Study Program, Postgraduate Program of the University of Islam Malang. Welcome you all to the campus of UNISMA.
As we know the theme of this first international ELTAR Conference is *Promoting Innovation and Transformation in English Language Teaching and Learning*. The selection of this theme must have undergone a very serious discussion and preparation as it will touch upon fundamental issues in today's agenda of improving the quality of English teaching and learning through innovative endeavors in the form of research on any possible aspects of teaching and learning activities. Any innovative ideas and research findings should be shared and disseminated to others so that they can be put into practice in the real classroom teaching-learning processes. It is at this context that the conference is carried out. I am fully convinced that this conference will provide us with an excellent opportunity to share and exchange ideas, knowledge, expertise, and experience as well as findings of research on the teaching of English in the Indonesian context and beyond. Moreover, I hope that this conference may also be a wonderful occasion for building and sustaining collaboration and networking among teachers of English and researchers of English language teaching around the globe.
Therefore, my special thanks should go to the invited keynote speakers, who have been willing to share their expertise to all of us in this conference. I would also like to congratulate the presenters, whose papers have been selected to be presented in this conference. To be chosen as presenters in this very prestigious academic forum must be a great achievement. I also would like to thank the participants, without whom the conference will never be a successful event. I hope you enjoy every moment of this wonderful conference.
Finally, I would like to offer my appreciation to the Director of the Postgraduate Program and the Head of English Language Education Study Program as well as all members of the organizing committee who have been working very hard to make the conference a great success. May Allah SWT bless you all. Thank you very much.

Wasslamu'alaikum War. Wab.

Prof. Dr. H. Maskuri, M.Si
Rector

PROCEEDINGS

INTERNATIONAL SEMINAR ON ENGLISH LANGUAGE TEACHING AND RESEARCH

TABLE OF CONTENTS

MINIMIZING RATERS' BIAS IN ASSESSING WRITING PERFORMANCE

Endah Yulia Rahayu
Indahr_99@yahoo.com
Universitas PGRI Adi Buana Surabaya

Abstract

Measuring writing performance, which is usually subjective can increase many interpretations which may not the same to the constructs. This problem is overcome with inter-rater reliability among raters to get an agreement which not always relating to the construct of the test and surely threaten their rating task unreliable. One of the threats of the construct validity is construct-irrelevant variance which includes dimensions beyond the construct and enables rater to measure the test taker's response too stringent or lenient. This is called bias because it distorts the test results and therefore the conclusion of the test taker's response based on scores is less valid. Rater bias can be caused by the halo effect that raters do not distinguish between different aspects of a composition. Another frequent bias is raters using middle of the scale. They are wary to use the two ends of the scale of the scoring rubric criteria. These errors are usually committed by novice raters. During rating task, they attempted to judge the criteria while that they are reading. However, the expert raters do not focus so much on certain things when as they read, but allow them to communicate with the texts on a more personal level. Then they evaluate the text more generally as a whole after they finish reading. Thus, they need to form a rater training to help novice developing a sense of the standards, as well as developing ways of approaching the rating task. Rater training also helps to improve intra-rater rather than inter-rater reliability and there is a correlation between training/experience and rating task. An Authentic rater training introduced by Kim (2016) which is characterized as a rater--centered bottom-up approach can be applied.

Keywords: raters'bias, test takers' response, rater training, rater severity

INTRODUCTION

The success or failure of writing performance assessment which requires test takers to perform real or actual task parallel to knowledge or skill being measured, is really typically determined by human raters. (Kane, Crooks, Cohen, 1996). In the assessment of second language writing today, they have to judge the modal form of the timed and impromptu writing test. The timed writing test is the test takers are given a fixed amount of time – usually 30 minutes to 60 minutes to write on a given topic provided while an impromptu test is the test takers are given a "prompt" providing a general introduction and context for their writing. Next, their written responses are read and scored by either one or more trained human raters or judges who judge their ratings based on some common criteria relating to the construct of the test. (Weigle, 2000).

Scoring of writing performance assessments which usually depends on the human raters, commits to the subjectivity of the scoring process and increase various factors not suitable with the construct. (Messics, 1996) This problem is overcome with the statistic calculation of inter-rater reliability (Durbar, Korets, Hoover, 1991) to obtain a more reliable score of increasing agreement. But the desirability of increasing agreement of

inter-rater reliability does not mean anything if what the raters agreeing on something are not in accordance with the construct. (Wiggle, 1999) It is obvious since raters of writing performance assessments have different personal and professional backgrounds. Thus what descriptor they actually appraise and what beliefs they possess during the rating task and not always the same and sometimes unclear. This surely threatens their rating task unreliable.

Relating to the language performance assessment, indeed validity and reliability is inseparable, that validity measuring to what is being assessed while reliability relating to how well what is being assessed. Just like in writing performance assessment, what is being measured in writing is also how it is being measured through writing (Bachman, 1990). In accordance with scoring validity, Shaw and Weir (2007) stated that scoring validity is standard and the tasks developed at the prompt are significantly valid in terms of contextual and cognitive parameters.

Thus, the problem is how to provide test takers with comparable treatment because they do not have a choice to the prompt they have to respond and the raters who read and rate their responses. The prompt for a test taker are usually taken from a prompt collection or more. Therefore it is not easy to imagine that any two prompts with the same or different topics will be totally comparable in every way. This can be a prompt effect, comparing the performances of a test taker who responses to a prompt and another test taker replying to another prompt. (Jennings, Fox, Graves, Shohamy, 1999) There can also be a rater effect, at the same way, that the test-takers' responses are scored by different raters who may have different severity and leniency in rating. Test takers are also possible to interact differently with different prompts. How comparable scores given by different raters to different test takers who respond different prompts will raise questions about their validity and reliability. It perhaps more importantly, raising questions of fairness. The fairness of an exam is free from any kind of bias which harms the quality of examinees' test, irrespective of race, religion, gender, or age. The test also need not to advantage any examinee or group of examinees, other than the examinee's lack of the knowledge and skills that the test is intended to measure. Therefore to conduct language performance assessment, teachers need to strengthen fundamental of fairness, validity, and reliability. (Kunnan, 2000)

VALIDITIY AND VALIDATING WRITING PERFORMANCE ASESSMENT

The primary concept of validity in language assessment is summed up by Robert Lado (1961) who states that a test needs to measure what is supposed to be measured, unless it is not valid". In a wider educational measurement domain, this concept emphasizes criterion validity, or the relation between measuring criteria and test scores. According to Messick (1989) in his seminal Educational Measurement, based on test modes of assessment, validity is a totally evaluative appraisal supported by empirical and theoretical evidence adequately and appropriately. Meanwhile, AERA, APA and NCME (1999) state the standards for educational and psychological testing encompass validity into content, criterion, and construct. The construct is defined as a concept of the test to measure the content and construct validity which are seen as key. Thus, the content and criterion validity cannot be appraised unless the construct is judged adequately. Thus, the different types of validity evidence support the constructs and forms of construct which are used in tests.

The evidential basis of test interpretation is construct validity. The construct needs to be adequately defined because no pretension can be made. After the construct is made

about the observed performances, the appropriateness of the pretensions needs to be assessed. (Cronbach, 1988). Meanwhile, Messick (1989) mentions two general threats to the construct validity: construct-irrelevant variance and construct underrepresentation. Constructs underrepresentation is defined as observations not including most important aspects of the construct. Meanwhile constructs-irrelevant observations include dimensions beyond the construct – a problem of appraising too much or little. Both constructs underrepresentation and construct-irrelevant variance arise alternative interpretations and arguments about what the test is measuring. In addition to the evidential basis of test is the claims and interpretations made to be meaningful and appropriate to the given particular context. The consequential aspects of validity cover social and cultural dimensions underlying constructs and with the societal consequences of utilizing tests. It relates to a definition of validity by making arguments about test's purpose and function which is always provisional and accumulate the evidence for particular interpretations and usage. The new evidence and observed consequences can support such interpretations. (Messicks, 1989)

RATER BIAS IN WRITING ASSESSMENT

The one of the threats of the construct validity is construct-irrelevant variance which includes dimensions beyond the construct and enables rater to measure the test taker's response too stringent of lenient. This is called bias because construct-irrelevant variance distorts the test results and therefore the conclusion of the test taker's response 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". (McManara, Roever, 2006) If responses of the test takers in the form of essay are scored differently from their equal ability, a construct-irrelevant variance affects the scores, causing test measures not only what it is intended to measure but something more, making the result an invalid source for interpretation. Biased tests harm all the educational and social institutions, since students might be admitted to a program or job for which they do not have the required ability and knowledge, while, on the other hand, qualified individuals might be rejected and deprived of their deserved positions and rights. (Saeidi, Yousefi, Baghayei, 2013)

It is an organized pattern of rater behavior that manifests itself in writing assessment. (Eckes, 2012) For example, raters may show unexpectedly high or low degrees of severity when scoring writing performance of test takers, or when using a particular scoring criterion to score writing performance of test takers. When raters show this kind of differential leniency or severity, they exhibit differential rater functioning. (Engelhard, G. , 2008) Kim (2009) also addressed the differential impact of rater background variables, in particular, rater language background on rater severity (Kim, 2009), and possible effects of rater training which causes rater severity (O'Sullivan et al, 2007).

Rater bias can occur in relation to various aspects based on the assessment situation and condition. In accordance with the prominent role of scoring rubric criteria in the complex process of assessing writing performance, bias can happen due to interaction between raters and the criteria. Wigglesworth (1993) states that some raters can score consistently or harshly on a criterion relating to grammar, fluency, vocabulary and whereas others may score more leniently on this criterion. Knoch et al. (2007) also studies the scoring criterion related to rater bias by concentrating on a comparison between face-to-face and online training procedures. They find out that in the training groups only some raters exhibit less bias after training, while the others even develop new biases.

In accordance with the rater behavior in the context of the Occupational English Test, McNamara (1996) notes raters' perceptions of grammar had a predominant influence on awarding test scores. This contrast remained the same even though raters were trained well, indicating the presence of a specific grammar-related bias that was difficult to change. In a study of rater types, Schaefer (2008) explores bias patterns of inexperienced raters of native English-speaker who evaluate EFL essays composed by the Japanese university students. The raters apply an analytic scoring rubric with six criteria: organization, content, style and quality of expression, mechanics, language use, and fluency. The MFRM Rater Criterion interaction analysis results a substantial proportion of significant bias terms. By sorting the flagged interactions into unexpectedly severe or unexpectedly lenient ratings, Schaefer managed to identify subgroups of raters sharing a particular rater bias pattern. For instance, if a subgroup of raters exhibits particular high severity toward content and/or organization, it also exhibited unusually high leniency toward language use and/or mechanics. Another subgroup can show a reverse bias pattern - high leniency toward content and/or organization but high severity toward language use and/or mechanics. A uniquely specific patterns of bias, indeed can be shared by subgroups of raters.

RATER TENDENCY AND CONSISTENCY

Engelhard (1994) states there are four major categories of rater errors. The first tendency is towards severity and the leniency which is a rater consistently giving lower or higher appraisal than a deserved performance. Engelhard (1994) mentions that it should be the best for raters to be continuum severity or leniency, if test takers are scored by raters who have much severity variance. Thus people can get higher scores than they deserve or some may get lower scores than they ought to have. This surely can affect the validity of the scoring. In addition to that, raters can differ in severity at different criteria in the scale. (Schaefer, 2008) For example, it may be relatively easier to get a score of four responses of test takers with one rater, but it is harder to get a score of six responses of test takers from raters. A solution to such problems is to provide the raters who are consistent in their severity. In addition to individual raters differing in severity, rating experience of language background has been studied since it affects the relative severity and leniency of groups of writing raters. (Wigle, 1999; Hill, 1996; Kondo-Brown, 2002)

The second tendency of rater error is halo effect, where raters do not recognize different aspects of composition. For example, some raters may form a general impression of the test takers' writing performance, after having seen a few writing responses and make subsequent judgments. The raters may simply stop paying attention to the written responses of test takers while others may (unconsciously) be tempted to make subsequent ratings consistent with earlier ratings. When a halo-effect occurs, the test takers may decrease their opportunities to demonstrate his or her writing proficiency which is a threat to the reliability of the examination. (Bechger, Maris, Hsiao, 2007)

The the third tendency is rater error that raters mostly use the middle range of the scale and are reluctant to use the two ends of the scale of the scoring rubric criteria. This creates no real sense of consistency to the overall raters' ratings. Meanwhile the fourth rater-error category is related to the third that the benchmarks and rangefinders at the criteria relating to the extent to which ratings are able to discriminate different test takers into different performance levels. If the raters fail to differentiate the written responses of the test takers based on the scoring criteria, then the purposes of measurement are

defeated. The four categories of the above errors are identified as cross-sectional that can happen at any raters at one particular time.

In addition to that, many assessments, however, are given many times to study rating consistency. Fitzpatrick et al (1998) conduct two investigations where exams of third, fifth, and eighth grade students of many of the subject areas are re-scored after one year and find out that the absolute standardized mean differences are relatively small - in the range of one-tenth up to two-tenths of a standard deviation. One of the exceptions is in Writing of Grade 5, where the mean difference can be considered large. They also calculate the correlations of total scores in the first and second sets of ratings. Correlations are consistently the highest in Mathematics, and consistently the lowest in Writing. Pearson correlations for third, fifth, and eighth grade writing are 0.58, 0.59, and 0.72, respectively. In this investigation, however, the raters in the first and second round are not the same people.

Another study about rating tendency is Cho (1999) observing ten raters to score the same 20 student essays four times, with an interval gap of four to six weeks between readings. This study finds high Kendall Taub correlation coefficient, with many raters reach internal consistency values higher than 0.7 across comparisons. Cho wonders whether there might be a possible memory effect, which presents a confound. Similar study is also conducted by Congdon and McQueen (2000). They study the scoring of 16 raters in seven rating sessions in nine days, where written performances rated on the first day are re-rated by the same raters on the last day. The raters read an average of 173 essays per day in order to make them not to memorize what scores they have given to essays they read more than once. On a daily basis, ratings for the writing responses of the test taker became more stable beginning with the fourth rating session. Congdon and McQueen find out a period of practice and getting used to the task was necessary for the raters who only have a half-day training session. The rating sessions are also divided by a weekend when there is no rating activity. The finding of a weekend effect suggests that re-training is needed for these raters. (Congdon, P. J., McQueen, J. , 2000)

THE EFFECT OF RATER TRAINING

Novice and experienced raters mostly differ, particularly in the way they rate. Huot (1993), in a think-aloud study, finds that when novice and experienced raters rating with the same holistic criteria, their reading essay of the test takers are quite different. In this case, the novice raters who are not given scoring guidelines, tend to make more comments as they started reading while the expert raters make more comments after they finish reading responses. The expert raters also make a greater percentage of personal comments because they already know what to evaluate in a composition and have a strategy for rating. Their strategies are not all the same, but each has a strategy that worked for them. (Huot, B. A. , 1993) In this regards Cumming, et al. (2002) also have the same finding. By comprehending the scoring criteria and possessing a strategy to score, expert raters do not to focus too much on particular texts as they read, and allow them to engage with the texts on a more personal level to evaluate the compositions more generally and as a whole after they read. In comparison, of novice raters, they are obviously attempting to develop judging criteria at the same time they are reading. Thus, they make more comments as they read along the text which at once result in more remarks having to do with the steps they were taking. It is not surprising that most novice raters having rating technique broke down at some point. It is obvious that the novice raters had the same criteria as the expert raters had. But their

attention is devoted to discovering the criteria and they have problem to engage with the texts in order to score more holistically. (Cumming et al, 2002)

Wolfe et al, (1998) confirm Huot's (1993) finding by using analitical rating scare. All the raters their studies are provided a rating scale and they are classified as competent, intermediate, and proficient. Like Huot, Wolfe et al find that proficient raters, comparing to the competent and intermediate raters, have fewer interruptions while reading and are able to judge until after they finish reading. They also make more general comments, and consider all textual features equally and use more rubric-related criteria. What these studies suggest, is that training might not have the same effect to all raters. Alternatively, the strategies of successful raters suggest the need of rater training that should help novice or inexperienced raters to emerge a sense of the criteria standards and to develop the most appropriate ways to approach the rating task.

Weigle (1998) observes eight experienced raters and eight inexperienced raters to score writing samples of a college placement test, in a pre- and post-training design. By applying multi-faceted Rasch methodology, Weigle finds that inexperienced raters are more severe and less consistent in their rating compared to experienced raters. Their training can reduce but cannot eliminate the differences in severity between the two groups of raters. The rating consistency of inexperienced raters, however, show much improvement after training. Rater training can help to improve intra-rater rather than inter-rater reliability, therefore there is a correlation between training/experience and rating task. Another study by Weigle (1999), using a pre- and post-training design, also invites the experienced and inexperienced raters rating essays of two different tasks: one task was calling on the test takers to make and defend an essay; the other task asked the test takers to interpret a graph into an essay. The results show that before training, the inexperienced raters are more severe in rating the paragraph essay. However, after training, this difference in severity disappeared. In accordance with think-aloud protocols, this study indicated that the two tasks elicited essays that were differently structured. The scoring rubrics were not as easy to use in scoring the graph essay for the inexperienced raters.

TRAINING FOR RATERS

A training for raters need to be conducted to minimize their bias during rating tasks because they can adopt and field-test the procedure in their own and also experience the test atmosphere. It is obvious because regarding to rating severity, both novice and experienced raters may behave more similarly when using analytic scales than they use holistic scales. In other words, raters' experience and possible raters' background impact ratings differently depending on the type of rating scales (Kim, C., 2016) because the purpose of rater training is to enhance the quality of raters' performance. (Weigle, S. C. , 1999)

Since one of the aims in major rater training involves monitoring rater behavior associated with rater-associated factors such as rating style or rating preferences, experience, and also how to the raters provide feedback accordingly to achieve the ultimate goal of enhancing inter-rater reliability (i.e. different raters scoring the same text). However, Weigle (1994) reports rater training does not always improve inter-rater reliability, but it can make raters more self-consistent. His statement apparently leads to a question to what factors contributing to develop rater self-consistency. There has not been much research that directly explores this question, but several studies report the positive effects of rater training to rater performance in many different aspects of the rating task.

According to Harsch and Martin's (2012), during the training, the raters should be engaged in a series of in-depth tasks of analyzing and revising the descriptors on the scale in order to reach the consensus about how to interpret different scripts with reference to scale description. Therefore rater training may require span over a two-month period to create an impressive and exemplary training. (Harsch, C., Martin, G. , 2012) However, such training may not always feasible in most real contexts. The reality of most rater training is likely to resemble the twohour norming session referred to as a typical rater calibration procedure. (Weigle, S. C., 1994)

Kim (2016) introduces an example of a rater training program which is characterized as a rater-centered and bottom-up approach which lasting less than two months. Her training protocol scaffolds to understand and apply the descriptors on the rating scale. This procedure surely activates the trainees' existing schemata and knowledge in rating. In this bottom-up training, the trainees can acquire new knowledge and technical concepts through a sequence of small tasks, than through top-down rater training which give abstract descriptors of raters. For many novice raters, in bottom up training, they can learn how to use a rating scale with predetermined descriptors includes an element of language acquisition. The procedure introduced by Kim resembles with language learning activities based on TBLT approach that language acquisition happens as a natural part of successful completion of communicative tasks. (Van den Branden, K. , 2006) In addition to that the Kim's training procedure provides guidance to translate descriptors into numerical scores in two ways: (1) focusing on the descriptors to match with numerical scores and (2) presenting numerical scores to match raters' judgments based on descriptors. Kim stated that her training procedure has realistic and practical advantages - easily adopting and adapting in most rater training contexts and rating scales, taking only 2~4 hours to complete the entire protocol and relatively inexpensive.

CONCLUSION AND SUGGESTION

Scoring of writing performance assessments requires the judgment of human raters who are usually subjective during the scoring process and can increase variance due to factors not related to the construct. This problem is overcome with inter-rater reliability to get agreement among raters. The agreement should related to the construct of the test. Different personal and professional background of raters affects their beliefs and predisposition when they rate the written responses of test takers. This threatens their rating task, unreliable.

A valid test should measure what it is supposed to measure, unless it is not valid. Validity is the level condition to which the evidence and theory support the interpretations of test appraisal. Three types of validity are content, criterion, and construct. The construct is the concept that a test is designed to measure the content and construct validity which are seen as a key. As a result, the content and criterion validity cannot be evaluated except by making reference to the construct. Two general threats to construct validity are construct underrepresentation and construct-irrelevant variance. Constructs underrepresentation is observations do not include all important dimensions of the construct. Meanwhile constructs-irrelevant observations include dimensions beyond the construct – a problem of measuring too little and a problem of measuring too much. Both threats can influence the scoring validity, where there can be alternate interpretations and arguments about what the test is measuring. Finally, consequential aspects of validity concerns with societal and cultural values underlying the constructs.

One of the threats of the construct validity is construct-irrelevant variance which includes dimensions beyond the construct and enables rater to measure the test taker's response too stringent of lenient. This is called bias because it distorts the test results and therefore the conclusion of the test taker's response based on scores is less valid. If responses of the test takers in the form of essay are scored differently from their equal ability, a construct-irrelevant variance affects the scores, causing test measures not only what it is intended to measure but something more, making the result an invalid source for interpretation. This bias is a organized pattern of rater behavior manifesting in a rating task which can be caused by raters' background of language and profession, inter-rater agreement, rater severity in scoring criteria, assessment situation and condition, and raters' inexperience in rating.

Next, raters can give consistent either lower or higher ratings than a real performance deserves. Based on the continuum of severity and leniency, test takers are scored by raters varying much in severity. Thus, some people can get higher scores than they should deserve or some should get lower scores than they ought to have, which could clearly affect the validity of these scores. Raters differ in severity and some studies investigate the aspects making relative severity and leniency of writing, such as rater experience. Another factor of rater error is the halo effect that the raters do not distinguish between different aspects of a composition. The most tendency which is mostly used is the middle of the scale. Indeed, most raters are reluctant to use the two ends of the scale of the scoring rubric criteria and this causes an artificial sense of consistency to the overall raters ratings.

The investigation of comparing novice and experienced raters compares the differences between them, particularly in the way they go about rating. Expert raters do not focus so much on particulars as they read but allow them to engage with the texts on a more personal level. Next, they evaluate the written responses more generally as a whole after they finish reading. However, novice raters are still apparently attempting to judge based on criteria at the same time that they are reading. The rationale is that they make more comments as they read along the text. Their effort to do many things at once results in more comments having to do with the steps they are taking. It is not surprising that most novice raters report that their rating technique break down at some points. It is obvious that the novice raters had the same criteria as the expert raters had, but their attention devote to discovering the criteria means. The novices also are not able to engage with the written responses and appraise them more holistically.

Training does not have the same effect to all raters and alternately the strategies of successful raters need to form a rater training that helps novice or inexperienced raters to develop a sense of the rating standards and task. Training can reduce but does not omit the differences in severity between the novice and experience raters. But, the consistency of inexperienced raters, show much improvement after training. Therefore, rater training suggests to enhance intra-rater rather than inter-rater reliability and there is an correlation between training/experience and rating task. Finally, an authentic rater training procedure which is characterized as a rater-center and bottom-up approach can be applied. The results indicate that before the training, inexperienced raters are more severe in scoring the essay. However, after training, this difference in severity disappeared. Based on several studies, training for raters lessens bias in rating writing performance, but others may develop new biases. Last but not least, training should be tailored wisely for any raters based on their needs, purpose and demands in order to shape the most effective raters.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Reseach Association.

Anderson, M., & Anderson, K. (2003). *Text Type in English 1-2.* Australia: Mc Millan Education.

Andrew Wright, D. B. (2006). *Games for Language Learning.* New York: Cambridge University Press.

Arikunto, S. (2006). *Prosedur Penelitian.* Jakarta: PT. Rineka Cipta.

Bachman, L.F. (1990). *Fundamental consideration in language testing.* Oxford: Oxford University Press.

Bechger, T,M., Maris, G., Hsiao, Y.P. (2007). *Assessing the size of halo-effects in performance-based tests and a practical colution to avoid halo-effects.* Cito, Arnhem: National Institute for Educational Measurement.

Braine, G., & May, C. (1996). *Writing from sources : a guide for ESL students.* USA: Mountain View, Calif. : Mayfield Pub. Co.,.

Burns, A. (2010). *Doing Action Research in English Language Teaching: a Guide for Practitioners.* New York: Routledge.

Cho, D. W. . (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Language Testing, 8*(1), 1-24.

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

Cronbach, L.J. (1988). Five perspective on the validity argument. In H. &. Wainer, *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cumming, A., Kantor, R., Powers, D. E. . (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Depdiknas. (2004). *Kurrikulum 2004 Standart Kompetensi Mata Pelajaran Bahasa Inggris Sekolah Menengah Pertama dan Madrasah Tsanawiyah.* Jakarta: Depdiknas.

Durbar, S.B., Korets, D.M., Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

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

Engelhard, G. . (2008). Differential rater functioning. Retrieved May 8, 2017, from https://www.rasch.org/rmt/rmt213f.htm

Englehard, G. (1994). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Finocchiaro. (1994). *English as a Second Language from Theory to Practise.* New York: Regent Publishing Company.

Fitzpatrick, A. R., Yen, W. M. . (1998). The psychometric characteristics of choice items. *Journal of Educational Measurement, 32*(3), 234-259.

Freeman, D. L. (2000). *Techniques and Principles in Language Teaching.* Oxford: Oxford University Press.

G Lewis, G. B. (1999). *Games for Children.* Oxford: Oxford University Press.

Harmer, J. (1993). *The Practice of English Language.* Essex: Longman Group UK Limited.

Harmer, J. (2006). *How to Teach English (new edition).* England: Pearson Education.

Harsch, C., Martin, G. . (2012). Adapting CEFdescriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228-250.

Haycraft, J. (1990). *An Introduction to English Language Teaching.* London: Longman.

Hill, K. . (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Language Testing*, 29-50.

Hongue, A. (2003). *The Essential of English, a Writer's Handbook.* New York: Longman.

Huot, B. A. . (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Huot, *Validating holistic scoring for writing assessment* (pp. 206-232). Cresskil, NJ: Hampton Press.

J Brewster, G. E. (1991). *The Primary English Teacher's Guide.* London: Penguin Books.

Jacobs H, e. a. (1983). *Testing ESL Composition: A Practical Approach.* Rowley, MA: Newbury House.

Jennings, M., Fox, J., Graves, B., Shohamy, E. (1999). The test-takers'choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456.

Kane, M.T, Crooks, T., Cohen, A. (1996). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kim, C. (2016). Rater training for scoring rubrics: Rater-centered and bottom-up approach . *MinneTESOL*.

Kim,Y.H. (2009). Aninvestigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187-217.

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.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Kunnan, A. (Ed.). (2000). Fairness and validation in language assessment. *19th Language Testing Reseach Colloquium, Orlando, Florida.* Cambridge: Cambridge University Press.

Lado, R. (1961). *Language Testing.* New york: McGraw-Hill.

M, W. J. (1982). *Teaching Vocabulary.* London: Biddles LTD.

Marianne, C. M. (2000). *Discourse and Context in Langugae Teaching. A Guide foa a Language Teacher.* New York: Cambridge University Press.

McManara, T, Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell Publishing.

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

Messicks, S. (1989). Valildity. In R. Linn, *Educational Measurement* (p. 13). New York: Macmillan.

Messics, S. (1996). Validity and wahsback in language testing. *Language testing*, 241-256.

Nasr, R. T. (1972). *Teaching and Learning English.* London: Longman group limited.

Neuman, W. L. (1991). *Social Research Methods: Qualitative and Quantitative Approaches.* Boston: Allyn and Bacon.

O'Malley, M., & Pierce, V. L. (1996). *Authentic Assessment for English Language Learners: Practical Approaches for Teacher.* New York: Addison Wesley.

O'Sullivan, B., Rignall, M. . (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. *Research in speaking and writing assessment* (pp. 446-478). Cambridge: Cambridge University Press.

Saeidi, M., Yousefi, M., Baghayei, P. (2013). Rater Bias in Assessing Iranian EFL Learners' Writing Performance. Iranian Journal of Applied Linguistics. *Iranian Journal of Applied Linguistics (IJAL), 16*(1), 145-176.

Schaefer, E. . (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 467-493.

Scott, T. (2002). *How to Teach Vocabulary.* Malaysia: Pearson Longman Education Press.

Scott, W. A., & Ytreberg, L. H. (1990). *Teaching English to Children.* London: Longman Inc.

Shaw, S.D., Weir, C.J. (2007). *Examining writing: Research and pratice in assessing second language writing.* Cambridge: Cambridge University Press.

Siahaan. (2008). Interactive Writing in the EFL class: the internet journal.

Sokano. (n.d.). Retrieved from www.sokano.com.

Sokano. (n.d.). Retrieved from www.sokano.com

Van den Branden, K. . (2006). *Task based language education: From theory to practice.* Cambridge, UK: Cambridge University Press.

Wallace, J. (1982). *Teaching Vocabulary.* London: Biddles LTD.

Weigle, S. C. . (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Weigle, S. C. (1994). Using FACETS to model rater training effects. *Language Testing Research Colloquium.* Washington, DC.

Weigle, S.C. (2000). *Assessing Writing.* Cambridge: Cambridge Univeristy Press.

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

Wigle, S.C. (1998). Using FACETS to model rater training effects. *Language Tesing, 15*(2), 263-87.

Wolfe, E. W., Kao, C. W., Ranney, M. . (1998). Cognitive differences in proficient and nonproficient essay scorers. . *Written Communication, 15*(4), 465-492.

Yolageldili, A. (2011). Effectiveness of Using Games in Teaching to Young Learners. *Elementary Education Online*, 219-229.