A STUDY ON INTRA-RATER RELIABILITY OF STUDENTS' SELF-ASSESSMENT OF THEIR ENGLISH WRITING

Ida Isnawati
IAIN Tulungagung, East Java

Abstract: The study is an attempt to estimate the intra-rater reliability of students' self-assessment of their writing performances and to find out whether rater training improves the reliability. The rater training used is adapted from the model developed by Herman, Aschbacher and Winters(1992). This quantitative study which employed equivalent time-samples design has two variables; they were the reliability coefficient of students' self-assessment and the rater training. The data were collected by asking 25 students conducting self-assessment on their four writing tasks and analyzed using Spearman coefficient correlation. It was found that the *rs* were 0.798, 0.772, 0.699 and 0.637; it showed that the consistency within students in assessing their own writing was moderately high. However, the *rs* of the intra-rater reliability of the self-assessment after the treatments is not higher than that of other experience being available in the absence of the treatment. It is concluded that rater training has not been able to improve the intra-rater reliability.

Key words: self-assessment, intra-rater reliability, rater training, writing

Assessment has a very important role in instructional design along with learning goal and learning activity. The three are considered as anchor points of instructional design (Djiwandono, 2008:2). In the past, assessment has always something to do with a paper-pen-test and such test has been recognized as the main tool of assessment. The role of test in classroom is very dominant especially in the1970s until some recent years. In the1970s, some techniques for testing grammar and the four skills were being well-developed (Richard, J.C., 2002: 32). Each language skill and component has its different technique of testing. The test should also fulfill some important principles of testing such as validity, reliability, and practicality in order to be considered as a good test.

In the current years, however, attention has shifted to alternative assessment, or, sometimes well known as authentic assessment. This kind of assessment gained its popularity because testing is said not to assess the full range of essential students' outcome nor capture important information about test takers' abilities (O'Malley and Pierce, 1996:2). Besides, authentic assessment is used not only to measure the students' ability but also to assist the students' learning.

Authentic assessment is different from the traditional one in that it actually asks students to show what they can do. Students are evaluated on

what they integrate and produce rather than on what they are able to recall and reproduce. This is in line with Gracia's and Pearson's conclusion that the main goal of alternative assessment is to gather evidence about how students are approaching, processing, and completing real life tasks in a particular domain (Macias, A. H. 2002: 339).

This kind of assessment is considered as innovation in classroom instruction because it has several important advantages over the use of a test (Macias, A. H. 2002: 339). Among the advantages of authentic assessment are (1) it does not intrude on regular classroom activities; (2) it absolutely reflects the curriculum; (3) because the data collected are based on the real-life tasks, it provides information on strengths as well as the weaknesses of a student; (4) it also provides a menu of possibilities, rather than any one single method for assessment; (5) it is more multiculturally sensitive and free of norm, linguistic, and cultural biases found in traditional testing.

One type of authentic assessment is self-assessment. Self-assessment can be defined as an appraisal made by a student of his or her own work on learning processes (O'Malley and Pierce, 1996:240). From this definition, it can be inferred that it is a new form of assessment which is definitely very different from what most people commonly think of. Generally, assessment is conducted by teachers, and students are the objects of assessment. Using self-assessment, however, the students are required to assess their own performance.

That is why a conventional view of language pedagogy might consider self-assessment to be an absurd reversal of the teaching learning process (Brown, H.D, 2001: 415). This is a common doubt raised by some people, "How could learners who are still in the process of acquisition be capable of interpreting an accurate assessment of their own performance?" Many teachers do not yet feel comfortable with it. In fact, "teachers do not believe in giving up this much control to students, whom they do not believe to be capable of self-assessment" (O'Malley and Pierce, 1996:36). They are concerned much with subjectivity, as not only professional teachers get difficulties in assessing productive language skills like writing but even so the students. Students may be either under estimate or overestimate themselves, or they may not have the necessary tools to make an accurate assessment. This is in line with the statement of Brown (2004:270), "…especially in the case of direct assessments of performance, they may not be able to discern their own errors".

However, research has shown a number of advantages of self-assessment: direct involvement of students in their own destiny, the encouragement of autonomy, and increased motivation because of self- involvement in the process of learning (Brown, H.D, 2004:270). By having self-assessment, students will take part actively in their own learning process because they have to reflect their own learning by finding their difficulties and problems as well as their strengths, and in turn, this will raise their awareness and responsibility. When they are responsible with their own learning, they will be able to set their own learning goal (O' Malley and Pierce, 1996) and have high motivation to reach the goal. At the same time, they will also do every effort to remove any obstacle inhibiting them from reaching their learning goal. In this state, they have been considered as autonomous learners.

Based on the various advantages self-assessment mentioned previously and the fact that self-assessment by the students is sometimes still doubted, it is crucial to improve the students' ability in assessing themselves and this can be done through rater training. Rater training is structured activities that may help students do self-assessment on their writing. When the students can do self-assessment well, they will be able to get the abovementioned advantages. On the contrary, if they do not, the learning goal will not be set appropriately. It may lead to the wrong learning direction.

One way of checking the students' ability in assessing themselves is by having intra rater reliability. According to Bachman (as quoted by Rini, 2011) intra rater reliability can be examined by having at least two independent ratings from the rater for each individual language sample. This is typically accomplished by rating the individual samples once and then re-rating them at a later time.

Some studies reported that the students have been successful in assessing themselves as objectively as their teacher's assessment. Bailey (1998) cited in Brown, 2004:270) conducted a study in which learners showed moderately high correlation (between .58 and .64) between self-rated oral production ability and scores on the OPI. It tells us that learners' self-assessments may be more accurate than we might suppose. Munoz and Alvarez (2007) conducted a study on correlation between student self-ratings and teacher ratings in oral language. From the study, it was concluded that it is possible for the students to self-assess with accuracy and they can show favorable attitudes toward this practice.

The abovementioned research mostly concern the inter-rater reliability by comparing the result of students' self-assessment with that of gotten from teacher rating or other criteria. Not many of them studied the intra-rater reliability examining the students' consistency in producing similar result when they have to rate twice. Besides, the students' writing has rarely been of the researcher's concern.

Based on the reasons described previously, this study was conducted to find out the average of two reliability coefficients of self-assessment in the absence of rater training, the average of two reliability coefficients of self-assessment with rater training treatment and whether the reliability coefficients of students' self-assessment with rater training are higher than those of without rater training.

RESEARCH METHOD

This study consisted of two stages (1) measuring intra-rater reliability of the self-assessment of English writing; and (2) finding the impact of rater training on the intra-rater reliability.

To measure the intra-rater reliability of the self-assessment of English writing, the students were asked to assess four different pieces of writing they have produced. Each piece of student's writing was assessed twice; the first one was directly after carrying out the writing task and the second one was done on the following day. It was based on Bachman's statement that "In the case of intra-rater reliability, it is needed to obtain repeated ratings from the same rater separated by a brief period of time" (Rini and Agustinus, 2009)

To find out whether rater training improves the intra-rater reliability, this quantitative study employed **Equivalent Time-Samples Design** suggested by Tuckman (1999:170). It is used when only a single group is available for study and the group requires a highly predetermined pattern of experience with the treatment (Tuckman, 1999:170). The design is diagrammed below.

X0 O1 X1 O2 X0 O3 X1 O4

O is observation of the Spearman coefficient correlation (rs) of the scores of direct self-assessment and the scores of self-assessment on the following meeting;

X1 represents treatment (rater training)

X0 is other experience being available in the absence of the treatment (without rater training)

In detail, the following description may explain the diagram.

X1 **Rater training is done in week 2.**

O2 Self-assessment of student's writing 2 (directly & on the following day). It was done in week 2

X0 Other experience being available in the absence of the treatment (without rater training). It was done in week 3.

O3 Self-assessment of student's writing 3 (directly & on the following day). It was conducted in week 3.

X1 **Repeated rater training is delivered in week 4.**

O4 Self-assessment of student's writing 4 (direct & on the following day). It was done in week 4.

There were two kinds of data in this study: (1) The reliability coefficient (rs) of students' self-assessment in the absence of the treatment (without rater training); and (2) The reliability coefficient (rs) of the students' self-assessment on the treatment (with rater training). All those data about the reliability coefficients of the students' self-assessment, either with rater training or without rater training, were gotten from two administrations of self-assessments to the students in each week. The first one was done directly after the students finished wring their narrative texts, and the second one was done in the next two days after the first administration.

The population of this study was the first-year students of English Education Department at IAIN Tulungagung in the academic year of 2013/2014. The sample was a class consisting of 25 students. Since it is possible that the school system does not allow intact classes to be divided to provide for random or equivalent samples, the study will apply purposive sampling (Ary et.al, 2010:156). The sample selection in this study was done purposively by taking a class which has a relatively good ability in English. This was so because the self-assessment activity requires the students to understand some instructions and questions in English.

There were two instruments used to collect the data in this study; they were (1) rating-scale for self-assessing the student's writing and (2) interview guide. Self-assessment sheet used in this study consists of 15 statements where six of

them deal with the purpose and organization of narrative text, five of them have something to do with word/sentence usage of narrative text, and the last 4 statements concern about the use of mechanic in writing.

To respond each statement, the students had to give a check in one of the four column choices. The first column is for "always", the second is "often", the third is "sometimes" and the last column is for "never". All those choices were then converted to score in which "always" represents 4, "often" represents 3, "sometimes" represents 2, and "never" represents 1.

Interview guide was used to collect additional information after the students did self-assessment. They were asked their opinion about the implementation of self-assessment in their writing class and their problems in doing so.

To analyze the data, a comparison of the average of *O1* and *O3* with the average of *O2* and *O4* will yield a result that is not likely to be invalidated by historical bias. The assumption is if the average of *O2, O4* is higher than the average of *O1*and *O3,* the rater training improves the intra-reliability of the self-assessment.

RESEARCH FINDINGS

Before the students assessed their own writing, they were given a writing task on the narrative text. For the first writing task, they were required to tell their own experience. Finishing writing the narrative text, the students assessed their own writing by filling out the self-assessment sheets. There were fifteen statements they had to respond related with the purpose/organization of the text, word/sentence usage and mechanics. See the Appendix for the detailed self-assessment sheet. On the following two days, they were shown their narrative text and were required to fill in the self-assessment sheet for the second time.

The next week the same procedures were done. However, before the students wrote their narrative text about a fairy tale, they were given rater training. The steps of rater training can be seen in the research method section. After they finished writing, they were directly asked to assess their own writing by filling out the self-assessment sheet. Two days later, the narrative texts they have written were given back to them and they were required to do self-assessment.

In the third week, the students were directly asked to write their narrative text about a fable and fill in the self-assessment sheets after they finished writing. The next two days, they were asked to fill in the self-assessment sheets for the second time as they were reading their narrative texts they have produced before.

In the fourth week, the rater training was done for the second time with the same procedures. Then, writing session was done before the students did their self-assessment. In the following two days, self-assessment was repeated after the students reread their narrative texts.

Table 1 shows the students' scores of their self-assessment in the first week until the fourth week when they had not got any rater training.

Table 1. The Result of the Students' Self-Assessment on Their Four Writing Tasks

No	Name	The First Week Observation (Without Rater Training)		The Second Week Self-Assessment (with Rater Training)		The Third Week Observation (Without Rater Training)		The Fourth Week Observation (With Rater Training)	
		1st Self-Assessment	2nd Self-Assessment	1st Self-Assessment	2nd Self-Assessment	1st Self-Assessment	2nd Self-Assessment	1st Self-Assessment	2nd Self-Assessment
1	LDA	52	52	60	59	60	59	52	60
2	LV	47	43	47	50	54	56	56	50
3	LF	41	35	58	58	55	54	54	55
4	LS	49	45	54	54	54	51	54	54
5	LZ	45	43	45	53	52	54	58	55
6	MAA	45	44	60	60	57	55	46	48
7	MSR	49	47	55	58	57	57	57	56
8	MFV	42	41	48	55	45	52	53	50
9	MQA	49	39	51	52	51	55	43	47
10	MC	37	42	47	52	55	56	55	56
11	NZ	51	58	60	60	60	59	60	59
12	NS	56	54	57	58	57	58	57	55
13	NS	50	49	57	55	54	52	57	50
14	NAS	57	57	53	58	54	50	59	59
15	NWS	50	43	46	46	46	46	50	48
16	NK	53	53	60	60	60	60	53	53
17	NUK	55	54	57	56	53	50	55	52
18	NUH	51	51	55	57	58	56	56	55
19	NML	43	45	51	50	51	51	55	54
20	NFW	46	42	59	58	53	56	54	52
21	NM	34	37	58	52	57	53	56	57
22	NM	38	37	53	39	55	53	37	44
23	NL	42	41	51	55	51	55	55	48
24	PN	56	48	56	57	56	59	55	55
25	WK	42	52	49	54	57	56	42	39

Table 1 shows that the students had various responses on their own writing performance. The lowest score was 34 indicating that in almost all statements they gave score 2 for their writing. The highest score was 60 and it shows that the student give perfect score (4) for all statements in self-assessment sheet. The table also shows that most students gave different scores for the same piece of writing they had produced.

From the data in Table 1, the correlation coefficient can be calculated for each week of observation. The results of the four observations are shown in Table 2.

Table 2. The intra-rater Reliability of the Self-assessment of the Writing Performances

Observation	The Availability of the Treatment	rs
Observation 1	No rater training	.798
Observation 2	Rater training	.772
Observation 3	No rater training	.699
Observation 4	Rater training	.637

Table 2 shows the results of observations -employing Spearman rank-order coefficient correlation- of the values of the intra-rater reliability of the self-assessment of the four writing performances. The first observation was done to seek the intra-rater reliability of the self-assessment of the first writing performance. The result articulates that the intra-rater reliability of the self-assessment of the first writing performance is 0.798. Second observation was done to measure the intra-rater reliability of the self-assessment of the second writing performance. They were done after the treatment. The calculation displays that the intra-rater reliability of the self-assessment of the second writing performance is 0.772. So were the third and the fourth observations done to seek the intra-rater reliability of the self-assessment of the third and the fourth writing performances respectively. The findings show that the intra-rater reliability of the self-assessment of the third and the fourth writing performances are 0.699 and 0.637 respectively.

From Table 2 it was also found that the *rs* were 0.798, 0.772, 0.699 and 0.637; it showed that the consistency within students in assessing their own speaking performance was moderately high. The number of the students who were able to judge their own writing consistently also increased.

In spite of the above fact, however, to find out whether the rater training has impact on the reliability it is needed to compare the average of two Spearman coefficient correlations (rs) on other experience being available in the absence of the treatment and the average of two Spearman coefficient correlations (rs) on the treatment. If the average of two Spearman coefficient correlations (rs) on the treatment is higher than the average of two Spearman coefficient correlations (rs) on other experience being available in the absence of the treatment, it can be said that the treatment has impact on the intra-rater reliability of the self-assessment of writing performance.

Table 3 shows the comparison of the average of *O1* and*, O3,* with the average of *O2,* and *O4.*

Table 5 The Comparison of the Average of *O1* and *O3* with that of *O2* and *O4*

	First Administration	Second Administration	Average
X1	*O2 (.772)*	*O4 (.637)*	***.705***
X0	*O1 (.798)*	*O3 (.699)*	***.749***

The result indicates that the average of *O2, O4,* (after treatments) is .705. It is lower than the average of *O1, O3,* (without treatment) that is .749. This

assumes that rater training has not been able to improve the intra-rater reliability of the self-assessment.

DISCUSSION

The result of the present study points out that in terms of reliability coefficients gotten from all four week observations, the students' consistency in rating their own writing turned out to decrease as they were introduced to rater training. However, when it comes to the analysis of the number of the students who made consistent judgment on their own writing, the result is different. When the students had to assess their writing twice, there were more students who were consistent when they got rater training than when they did not get rater training.

These contradictory facts between the decreasing reliability coefficient when rater training was given and the increasing number of students who judged consistently when rater training was given may be explained this way. It was true that the number of the students who self-assess their writing more consistently increased as rater training was given. However, some students who judge inconsistently in their self-assessments made extreme score differences when rater training was provided. As one proof, a student with initial name NM made 14 score differences in week 2 when rater training was given.

This extreme score and other big score differences definitely decreased the reliability coefficient of students' self-assessment when rater training was provided although there were more students who could rate their own writing consistently after rater training was given.

Regardless of the explanation above, the result of calculation of reliability coefficient from those sets of self-assessment scores cannot be ignored at all. The result of reliability coefficient calculation is still used to determine the final result of this present study.

Thus, based on the result of reliability coefficient calculation, the finding of the present study definitely contradicts with the previous study done by Rini (2011) showing that the students' consistency in assessing their speaking performance improved as the rater training was given to them.

This fact may be caused by some factors. One of them is the students' familiarity of such assessment procedures. For the students involved in this study, this is their first experience to do self-assessment. They find it quite uncomfortable initially to have such new way of assessment since they are accustomed to be assessed by their teachers as in traditional education system (Brown, 2004:277).

The students also do not really understand the advantages of having self-assessment and unfortunately, the researcher did not get plenty of time to make them sure about the benefits they will get when they are able to assess their own writing. According to Brown (2004:277), it is not enough to simply toss a self-checklist at students and then walk away. Systematic follow-up can be accomplished through further self-analysis, journal reflection, written feedback from the teacher, conferencing with the teacher, purposeful goal-setting by the student, or any combination of the above. From this, we know that one or two

weeks will not be enough to train the students to be able to do self-assessment well.

Besides, the students got difficulty in defining some of writing traits. Brown (2004:21) states that rater reliability is particularly hard to achieve since writing proficiency involve numerous traits that are difficult to define. This is true with the students in this study. Many of them got difficulty in understanding the terms used in self-assessment sheet.

However, despite the decreasing trend of the students' intra rater reliability from week to week, their consistency in doing self-assessment on a piece of writing they have written was moderately high to high. According to Weigle (2002:135), the closer the coefficient to 1, the stronger the reliability is. The coefficient correlations of .798, .772, 0.699, and .637 are closer to 1. This finding suggests that it is possible to involve the students in the process of assessment of their own learning. The notion of traditional assessment that self-assessment is an absurd reversal of the teaching learning process (Brown, H.D, 2001: 415) should be denied. This is also the answer for the common doubt raised by some people of the students' capability of having objective assessment (O'Malley and Pierce, 1996:36). When the objectivity is concerned, it cannot be guaranteed also that teacher is more objective than the students in doing assessment. Experience and awareness are badly needed to be able to do assessment more objectively.

CONCLUSIONS AND SUGGESTIONS

The results of the observations show that the intra-rater reliability of the self-assessment of English speaking performance is moderately high; the rs is between 0.637 and 0.798. The students' consistency in self-assessing their speaking performance was moderately high to high. The comparison of the average of the two rs values of the intra-rater reliability of self-assessment of writing performance on other experience being available in the absence of the treatment (0.749) and the average of the two rs values of the intra-rater reliability of self-assessment of writing performance on the treatment (0.705) proves that rater training has not been able to improve the intra-rater reliability of self-assessment of writing performance.

However, this study also concludes that the intra-rater reliability of the self-assessment of the writing performance is moderately high; the r_s is between 0.637 and 0.798. The number of the students who were able to assess their writing consistently also increased after the rater training was given. It means that the students were consistent enough in self-assessing their writing performance and this definitely can be improved by giving more intensive rater training.

Therefore, some suggestions need to be addressed to the teachers to introduce self-assessment with rater training to the students and implement it in the classroom. The students are also suggested to implement self-assessment more independently because they will be more aware of their own learning, especially in learning to write well. Finally, it is recommended for the future researcher to have more intensive rater training for the students with clear

criteria and numerous benchmark papers. Besides, the language used in self-assessment should be understood well by the students. So, the use of native language is recommended. The future researchers should also be able to manage the students' boredom in doing self-assessment by giving more various and challenging writing tasks. Finally, it is crucial to give them longer time to self-assess their writing. This will give them chance to think clearly about their strengths and weaknesses.

REFERENCES

Ary, D., Jacobs, L.C., & Razavieh, A. 2010. *Introduction to* Research *in Education*. Wadsworth: Cengage Learning

Brown, H. D. 2001. *Teaching by Principles: An Interactive Approach to Language Pedagogy*. White Plains, NY: Addison Wesley Longman.

Brown, H.D. 2004. *Language Assessment: Principles and Classroom Practices.* White Plain, New York: Pearson Education

Cohen, A.D. 1994. *Assesing Language Ability in the Classroom (2nd Edition).* Wadsworth: Heinle & Heinle Publishers

Djiwandono, M. S. 2008. Tes Bahasa. Jakarta: Indeks

Hughes, A. 1989. Testing for Language Teachers. Cambridge: Cambridge University Press

Macias, Ana H. 2002. *Alternative Assessment: Responses to Commonly Asked Questions*. In Richards, Jack C. & Renandya, Willy A. 2002. *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: Cambridge University Press, page 339

Munoz, A & Alfarez, M. 2007. Students' Objectivity and Perception of Self Assessment in an EFL Classroom. The Journal of ASIA TEFL Vol. 4 No. 2, pp. 1-25. Summer 2007

O'Malley, J. M. & Pierce, V. P.1996. *Authentic Assessment for English Language Learners.* Addison-Wesley Publishing Company

Richards, J. C. 2002. 30 Years of TEFL/TESL: A personal Reflection. The 50th TEFLIN International Conference, Surabaya. 29,30, 31 October 2002. page 32

Rini, N. 2011. Improving the Intrarater Reliability of Self-Assessment of English Speaking Performance. *English Education Journal. Vol. 1 No.1*.

Tuckman, B.W. 1999. *Conducting Educational Research (5th Edition)* Orlando: Harcourt Brace & Company.

Weigle, S. C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press