THE EFFECTIVENESS OF RATER TRAINING IN IMPROVING THE SELF-ASSESSMENT INTRA-RATER RELIABILITY OF ENGLISH SPEAKING PERFORMANCE

Nur Rini

Abstract

This quantitative study is an attempt to estimate the intra-rater reliability of student self-assessment of their speaking performances and to find out whether there is significant difference between the self-assessment intra-rater reliability of speaking performance without training and that of with training. The rater training used is adapted from the model developed by Herman, Aschbacher and Winters (1992). This study which employed equivalent time-samples design collected data by asking 45 students to conduct self-assessment on their six speaking performances. It was found that the range of r_s was 0.611 to 0.752 which means the consistency within students in assessing their own speaking performance was moderate high to high. The intra-rater reliability of the self assessment after the treatments is higher than that of other experience being available in the absence of the treatment. It is concluded that rater training improves the intra-rater reliability. Therefore, it is suggested to train the students on how to assess before employing self-assessment in speaking instructions.

Key words: *self-assessment, intra-rater reliability, rater training, speaking*

Introduction

Self-assessment is appraisal by a student of his or her own work or learning processes (O'Malley and Valdez Pierce 1996; Sawyer, Watson, and Adams 1989). This study defines self-assessment as an activity in which the student is asked to describe his or her speaking performance by filling the self-assessment form which is designed by the teacher and the students. To estimate the intra-rater reliability of the self-assessment the student is asked to self-assess every of his or her own speaking performance twice. The first self-assessment is done directly after the speaking task is completed and the second is done the following day while he or she is watching the video recording of his or her speaking performance. In this study, the intra-rater reliability means the consistency within the students in self-assessing their English speaking performance. It is assumed

that rater training is needed to improve the reliability. Rater training is structured activities that this study believes may help students do self-assessment on their speaking performance.

According to Brown (2004:251), self-assessment is one of the forms of alternatives in assessment. It is to collect additional measures of students in an effort to triangulate data about students in minimizing the weaknesses of standardized tests. Self assessment offers certain benefits: direct involvement of students in their own destiny, the encouragement of autonomy, and increased motivation because of their self-involvement. For this reason the students are to have self-assessment skill.

Furthermore, self-assessment is providing opportunity for students to do reflecting of their learning (O'Malley and Pierce 1996; Johnston 1987; Carroll and Hall 1985). Self-assessment is the basis for setting the individual learning goal (O'Malley and Pierce 1996). Since setting the learning goal is based on the self-assessment, the intra-rater reliability is required. If the students are not able to do self-assessment well, the learning goal will not be set appropriately. It may lead to the wrong learning direction. The need for improving the intra-rater reliability in doing self-assessment is crucial.

Regardless of the advantages of employing student self-assessment, 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 Valdez Pierce, 1996:36). They are concerned much with subjectivity, as not only professional teachers find difficulties in assessing productive language skills like speaking but even so the students. Speaking is "a skill which deserves attention every bit as much as literary skills" (Bygate 1987:vii). Students may be either underestimate or overestimate themselves, or they may not have the necessary tools to make an accurate assessment. Furthermore, "especially in the case of direct assessments of performance, they may not be able to discern their own errors" (Brown, H.D. 2004:270). In contrast, Bailey (1998) cited in Brown, H.D. (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.

There are some objections, which are usually due to the technical difficulties, that teachers do not want to employ self-assessment in speaking instructions. However, a recent study done by Liu Qin and Wang Li (2008) on a portfolio approach - which is self-assessment is considered as a crucial part of it - to oral English assessment found that it was workable and could yield reliable results to assess students' oral English proficiency. Moreover, the students participating in the experimental group showed that they were very happy to be able to reflect upon and assess their own progress and had more confidence in improving oral English proficiency.

Hadley and Mort (1999), de Wet, der Walt and Niesler (2009) assume that there is a tight correlation between scoring system and rater reliability and they suggest that the scoring bands and their meaning need to be made explicit. Two observations on the use of self-assessment have been done, i.e. that on writing by Hall and that on oral English by Burn (both in Johnston, 1987: 125). The studies suggested teacher to let students explore the elements used for building good presentations and make the elements as the criteria for assessing their own work. The process will bring the students to understand the notion of setting their own learning goals.

The study on inter-rater reliability done by Bailey (1998. cited in Brown, H.D. 2004:270) showed moderately high correlation (between .58 and .64) between self rated oral production ability and scores on the OPI. Meanwhile, the studies on the intra-rater reliability have rarely been conducted. Therefore, the study on improving the intra-rater reliability in the self-assessment of speaking performance through rater-training is worth conducted.

"Intra-rater reliability is the consistency within raters (Bachman, 2004: 169)." In test scores that are obtained subjectively such as ratings of oral presentations, a source of error is inconsistency in these ratings. In the case of single rater, the concern with the consistency within that individual's ratings or with intra-rater reliability is required. Further description about intra-rater reliability is quoted from Bachman (1990:178-180):

> When an individual judges or rates the adequacy of a given sample of language performance, whether it is written or spoken, that judgment will be based on a set of criteria of what constitutes an 'adequate' performance. If the rater applies the same set of criteria consistently in rating the language performance of different individuals this will yield a reliable set of ratings.

In addition, Bachman (1990:178-180) recommends two ways to measure the intra-rater reliability of ratings, using Spearman rank-order coefficient or coefficient alpha.

> ...To examine the intra-rater reliability of ratings, at least two independent ratings from the rater for each individual language sample are needed. This is typically accomplished by rating the individual samples once and then re-rating them at a later time (in different, random order). Once the two sets of ratings have been obtained, the reliability between them can be estimated in two ways. One way is to treat the two sets of ratings as scores from parallel tests and compute the Spearman rank-order coefficient between the two sets of ratings, interpreting this as an estimate of reliability. ...

The intra-rater reliability of the self-assessment of English speaking performance can be estimated if there is a students group of speaking instructions employs self-assessment. Conducting self-assessment on speaking performance calls for speaking rating scale. It is suggested that the elements included in the rating scale are the results of teacher-student agreement (O'Malley and Pierce, 1996).

The purposes of the study are to measure the intra-rater reliability of the self-assessment of English speaking performance; and to find out whether there is significant difference between the self-assessment intra-rater reliability of speaking performance without training and that of with training.

Method of Investigation

The study consisted of three stages: (1) questionnaire survey on oral rating scale to set the rating scale; (2) measuring intra-rater reliability of the self-assessment of English speaking performance; and (3) finding the impact of rater training on the intra-rater reliability.

A questionnaire survey was conducted to collect the students' opinions on actual rating scales (Please refer to Appendix I for the questionnaire.). It goes with Aschbacher's argument that "One of the characteristics of performance assessment is that the criteria are made public and known in advance" (1991, in O'Malley and Pierce, 1996:

O_3 Self-assessment of speaking performance 3 (direct & recorded): Telling someone's personality. It was conducted in week 9.

X_1 **Repeated rater training 1 was delivered in week 11.**

O_4 Self-assessment of speaking performance 4 (direct & recorded): Telling my job preferences. It was done in week 12.

X_0 Other experience being available in the absence of the treatment

O_5 Self-assessment of speaking performance 5 (direct & recorded): Text retelling. It was accomplished in week 14.

X_1 **Repeated rater training 2 was through in week 16.**

O_6 Self-assessment of speaking performance 6 (direct & recorded): Describing a country. It was carried out in week 17.

As it is stated by Tuckman (1978:139) that "it is a form of time-series design but, rather than introducing the treatment (X_1) only a single time, it is introduced and reintroduced, with some other experience (X_0) being available in the absence of the treatment."

The subjects of the study were 45 first-year students of D3 Program of Business Administration Department of Semarang State Polytechnic, academic year 2008/2009. The number has met the minimum requirement for doing the correlation analysis as it goes with what Mantra (in Singarimbun, M. and Effendi S. (Eds). 1989) states that if correlation technique is used in the analysis the minimal sample of 30 should be fulfilled. There were three instruments to conduct self-assessment and collect the data: (1) Questionnaire on oral rating scales which was used to set the rating scale to assess students' speaking performance (See Appendix 1); (2) Six speaking tasks that were adapted from T*he New Interchange Book 2 (*Richards, Hull, and Proctor 1997): They were used to conduct the 3-minutes oral presentation; and (3) self-assessment form and the scoring guide (See Appendix 2 and 3).

If the value of r_s reaches 0.7, it can be concluded that the intra-reliability of self-assessment is considered high. It goes with Lado's statement (1961) cited in Hughes (2003: 39) that "oral production test may be in .70 to .79 range." He adds that "a

reliability coefficient of .85 might be considered high for an oral production test but low for a reading test."

A comparison of the average of O_1, O_3, and O_5 with the average of O_2, O_4, and O_6 will yield a result that is not likely to be invalidated by historical bias (Tuckman, 1978:140). The assumption is if the average of O_2, O_4, O_6 is higher than the average of O_1, O_3, O_5, the rater training likely improves the intra-reliability of the self-assessment.

Table 1 The Analysis Design to Find out the Impact of the Rater Training on the Intra-rater Reliability of the Self-assessment of Speaking Performance

	First Administration	Second Administration	Third Administration
X_1	O_2	O_4	O_6
X_0	O_1	O_3	O_5

To guard the external validity, which is one weakness of employing the design, the study held three administrations instead of having two administrations. If the effect of X_1 of the third administration is the same as its effect when introduced and reintroduced, then it would make valid conclusions about the continuous of X_1 from a study using the equivalent time-samples design (Tuckman 1978).

The rater training used was adapted from the model developed by Herman, Aschbacher and Winters (1992 as cited in O'Malley and Valdez Pierce, 1996). There are five phases described below.

(1) O*rientation to the assessment task*
- introducing the purposes of the assessment;
- describing who will use the assessment results;
- discussing the objective being assessed;
- describing the prompts and student directions;
- giving an overview of the scoring rubric; and
- taking the assessment themselves so they understand the mental processes that are being called on as they take the assessment.

(2) Clarification of the scoring rubric
- Discussing the scoring rubric and its components in small groups;
- Thinking back on the mental processes that are called for in responding to the prompt and how the rubric taps into these processes; and

- Reviewing the anchor performances (representative products or performances used to characterize each point on a scoring rubric or scale.

(3) Practice scoring
- Scoring a speaking performance in groups and individually and discussing the rates;
- Taking notes while scoring, providing reasons why they assigned the scores; and
- Attempting for establishing consensus in their ratings.

(4) Check reliability: Comparing the students awarded scores with the teacher awarded scores to check the reliability.

(5) Follow-up: The students were invited to do reflection on how they assessed the recorded speaking performances, to find out whether they overestimate or underestimate or they give about the same score awarded by the teacher.

The limitation of the study is that the student-teacher and student-peers discussions on the recorded speaking performance before the students did the second self-assessment might cause bias in this study.

Findings

The Rating Scale for the Self-assessment of Speaking Performance

The analysis of the core component of the questionnaire showed the first seven items among nine elements which have the highest mean scores. They were taken to put in the rating scale. They are grammar, pronunciation, loudness, vocabulary, body language, task and fluency. Two elements – cohesion and strategy - were excluded as they had the lowest means, 2.87500 and 3.06250 respectively. From the class discussion with the students, it was revealed that they felt that the two elements were too difficult to understand. There were only few answers to the open-ended question, and the answers were the clarification of the seven elements.

Many previous studies did not include the element of loudness. Loudness is the volume of the speaker's voice. This study considers the element is important to improve since the fact that not only most female but also male subjects of the study do not speak loudly when they do presentation in front of the class, even the teacher who sits about one meter from the student can hardly catch the students' words.

A class discussion was held to assign the weight on every element. Table 2 shows the result of the analysis on the questionnaire responses and the class discussion.

Table 2. The Core Component of the Questionnaire and the Weight of the Element

ITEM	ELEMENT	MEAN	NOTES	WEIGHT
1	Task	3.56250	Chosen	2
2	Pronunciation	4.35417	Chosen	2
3	Vocabulary	3.93750	Chosen	1
4	Grammar	4.43750	Chosen	2
5	Fluency	3.33333	Chosen	1
6	Cohesion	2.87500	Not chosen	-
7	Strategy	3.06250	Not chosen	-
8	Body Language	3.72917	Chosen	1
9	Loudness	4.06250	Chosen	1

With reference to the analysis result, the seven elements were used in the self-assessment (Please refer to Appendix 2: Self-assessment Form).

Intra-rater Reliability

The scores awarded by the students were entered into the Excel worksheet and calculated to get the final scores. Based on the agreement on the given weight (please see Table 2) the equation to calculate the final score is (2XTask score + 2XPronunciation score + Vocabulary score + 2XGrammar score + Fluency score + Body Language score + Loudness score)/10.

For example, if a student circles the highlighted numbers shown in the following table, the final score is (2X4+2X5+4+2X3+6+7)/10=41/10=**4.1.** The final scores were used as the data to observe the intra-rater reliability.

Table 3. Example of Student's Awarded Score

Item	Statement	not very well ←----→very well	Notes

1	I can complete the task	1	2	3	4	5	6	7	
2	I can pronounce the words	1	2	3	4	5	6	7	
3	I use appropriate vocabularies	1	2	3	4	5	6	7	
4	I use appropriate grammar	1	2	3	4	5	6	7	
5	I keep my presentation fluent	1	2	3	4	5	6	7	
6	I use eye contact, facial expression, and gestures to help convey my ideas.	1	2	3	4	5	6	7	
7	I can make my presentation audible.	1	2	3	4	5	6	7	

The results of the six observations - employing Spearman rank-order coefficient correlation - of the values of the intra-rater reliability of the self-assessment of the six speaking performances are presented in the Table 4.

Table 4. The Intra-rater Reliability of the Self-assessment of the Speaking Performances

Observation	The Availability of the Treatment	r_s
Observation 1	No rater training	.611
Observation 2	Rater training	.634
Observation 3	No rater training	.652
Observation 4	Rater training	.658
Observation 5	No rater training	.682
Observation 6	Rater training	.752

First observation was done to seek the intra-rater reliability of the self-assessment of the first speaking performances. The self-assessments were done on other experience being available in the absence of the treatment. The result articulates that the intra-rater reliability of the self-assessment of the first speaking performance is 0.611. Second

observation was completed to measure the intra-rater reliability of the self-assessment of the second speaking performance. They were done after the treatment. The calculation displays that the intra-rater reliability of the self-assessment of the second speaking performance is 0.634. So were third, fourth, fifth and sixth observations done to seek the intra-rater reliability of the self-assessment of the third, fourth, fifth and sixth speaking performances respectively. The findings show that the intra-rater reliability of the self-assessment of the third, fourth, fifth and sixth speaking performance are 0.652, 0.658, 0.682, and 0.752 respectively.

Impact of Rater-training on the Intra-rater Reliability

To find out whether there is significant difference between the self-assessment intra-rater reliability of speaking performance without training and that of with training, it is needed to compare the average of three Spearman coefficient correlations (r_s) on other experience being available in the absence of the treatment and the average of three Spearman coefficient correlations (r_s) on the treatment. If the average of three Spearman coefficient correlations (r_s) on the treatment is higher than the average of three Spearman coefficient correlations (r_s) 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 speaking performance.

The following table shows the comparison of the average of O_1, O_3, and O_5 with the average of O_2, O_4,and O_6.

Table 5 The Comparison of the Averages of O_1, O_3, and O_5 with of O_2, O_4,and O_6

	First Administration	Second Administration	Third Administration	Average
X_1	O_2 (0.634)	O_4 (0.658)	O_6 (0.752)	0.681
X_0	O_1 (0.611)	O_3(0.652)	O_5 (0.682)	0.648

The result indicates that the average of O_2, O_4, O_6 (after treatments) is 0.681, it is higher than the average of O_1, O_3, O_5, (without treatment) that is 0.648. The assumption

has been proved right, that if the average of O_2, O_4, O_6 is higher than the average of O_1, O_3, O_5, the rater training improves the intra-rater reliability of the self-assessment.

DISCUSSION

It is possible to quantify the reliability of a test in the form of a reliability coefficient. Lado's statement (1961) on reliability coefficient of good oral production test as cited in Hughes (2003:39) can be used as a help in judging whether the reliability is considered high or low, that "oral production tests may be in the .70 to .79 range." he adds that "a reliability coefficient of .85 might be considered high for an oral production test but low for a reading test."

It has been mentioned in the findings that the results of the observations are as follows: $O_1 = 0.611$; $O_2 = 0.634$; $O_3 = 0.652$; $O_4 = 0.658$; $O_5 = 0.682$; and $O_6 = 0.752$. In relation to the range of intra-rater reliability of the self-assessment of the speaking performance is 0.611 to 0.752; it can be said that intra-rater reliability of the self-assessment of the speaking performance is moderate high to high.

Many teachers do not yet feel comfortable with assigning students to do assessment; in fact O'Malley and Valdez Pierce (1996:36) argue that, "teachers do not believe in giving up this much control to students, whom they do not believe to be capable of self-assessment." They are concerned much with subjectivity. Also Brown says (2004:270) that, "especially in the case of direct assessments of performance, they may not be able to discern their own errors." In contrast, 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; and it is seconded by the finding of this study that the range of intra-rater reliability of the self-assessment of the speaking performance is 0.611 to 0.752. The students' consistency in self-assessing their own speaking performance is good. In this respect, teachers may learn to give more control to students to do self-assessment.

The average of three Spearman coefficient correlations (r_s) on the treatment (0.681) is higher than the average of three Spearman coefficient correlations (r_s) on other experience being available in the absence of the treatment (0. 648), it can be said that the

treatment has impact on the intra-rater reliability of the self-assessment of speaking performance. The figure 1 yields the comparison.

Although there are increases of the coefficient correlations (r_s) on other experience being available in the absence of the treatment, i.e. an increase of 0.041 of first X_0 to second X_0 (0.611 to 0.652) and an increase of 0.03 of second X_0 to third X_0 (0.652 to 0.682), the increases are lower than the increases of the coefficient correlations (r_s) on the treatment i.e. an increase of 0.024 of first X_1 to second X_1 (0.634 to 0.658) and an increase of 0.094 of second X_1 to third X_1 (0.658 to 0.752). The increases of the coefficient correlations (r_s) on other experience being available in the absence of the treatment may happen due to maturation.

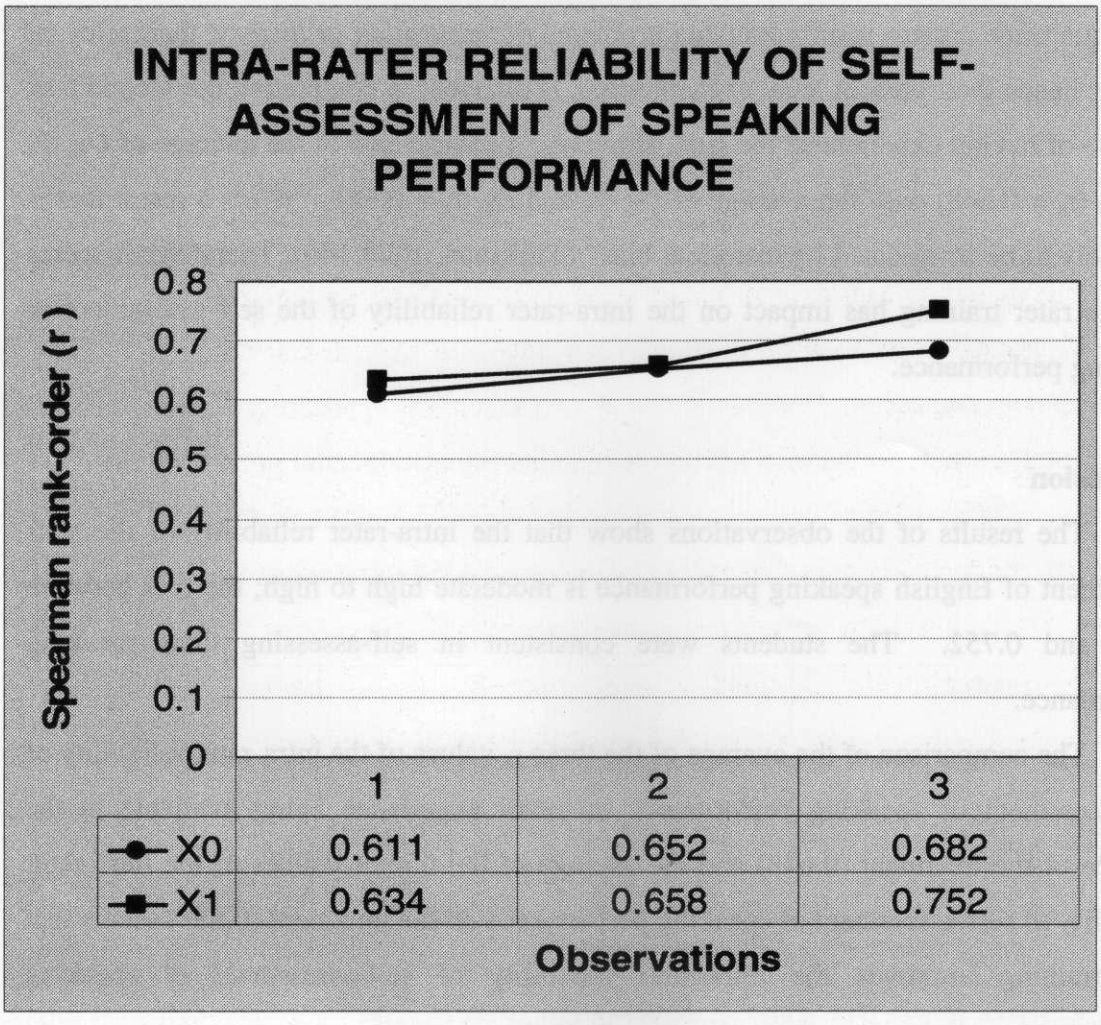

Figure 1. The Intra-rater Reliability of the Self-assessment of Speaking Performance on the Other Experience Being Available in the Absence of the Treatment and of on the Treatment

Although all subjects serve as both the experimental and the control group, which the selection variables can usually be considered to be adequately controlled; however, there are many situations where this technique cannot be used because the experimental experience will have an effect on a person's performance in the control activity or vice versa (Tuckman, 1978:106). In this study, for instance, after being experienced self-assessing their speaking performance and having the rater training, the subjects will no longer be naïve and performance on the control task will reflect the subject's experience on the training. In other words, while controlling adequately for selection bias this technique often creates insurmountable problems of *maturation* or *history*; their relevant history, hence their present level of maturation, is different in completing the second task because of having experienced the first. However, a comparison of the average of O_1, O_3 and O_5 (r_s = 0.648) with the average of O_2, O_4 and O_6 (r_s = 0.681) "yields a result that is not likely to be invalidated by historical bias" (Tuckman, 1978:140). Therefore, it seems that the rater training has impact on the intra-rater reliability of the self-assessment of speaking performance.

Conclusion

The results of the observations show that the intra-rater reliability of the self-assessment of English speaking performance is moderate high to high; the r_s is between 0.611 and 0.752. The students were consistent in self-assessing their speaking performance.

The comparison of the average of the three r_s values of the intra-rater reliability of self-assessment of speaking performance on other experience being available in the absence of the treatment (0.648) and the average of the three r_s values of the intra-rater reliability of self-assessment of speaking performance on the treatment (0.681) seems that rater training improves the intra-rater reliability of self-assessment of speaking performance.

Finding that the student's intra-rater reliability of the self-assessment of English speaking performance is moderate high, the teachers are suggested to learn to give their students more control in doing self-assessment.

Improving the intra-rater reliability in doing self-assessment is crucial. If the students are not able to do self-assessment well -underestimate or overestimate their speaking skill- the learning goal will not be set appropriately. It may lead to the wrong learning direction. It is suggested to hold rater training before employing self-assessment since the study proves that the student's intra-rater reliability of the self-assessment of English speaking performance is likely improved by conducting rater training.

Due to the limitation of this study, if a similar study is conducted, it is suggested to ask the students to do the second self-assessment before conducting the student-teacher and student-peers discussions on the recorded speaking performance.

References

Bachman, L. 1990. *Fundamental Considerations in LanguageTesting.* Oxford: Oxford University Press.

Bachman, L. 2004. *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

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

Bygate, M. 1987. *Speaking.* Chandin, C.N. & H.G. Widdowson (Eds). New York: Oxford University Press.

Carroll, B.J. and P.J. Hall. 1985. *Make your Own Language Tests.* Oxford: Pergamon Press Limited.

de Wet, F., Van der Walt, C., and Niesler, T.R., 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication,* Volume 51, Issue 10, October 2009, Pages 864-87. Available at http://www.sciencedirect.com [accessed 10/20/09]

Hadley, G. and Mort, J. 1999. An Investigation of Interrater Reliability in Oral Testing. .On line at

http://www.informaworld.com/smpp/content~db=all~content=a713696212 [accessed 01/20/09]

Hughes, A. 2003. *Testing for Language Teachers.* (2nd Ed.). New York: Cambridge University Press.

Johnston, B. 1987. Assessing English: Helping Students to Reflect on Their Work. Philadelphia: Open University Press.

Liu, Q., and Li, W. 2008. A Portfolio Approach to Oral English Assessment for English Majors in China. 6th Asia TEFL. 2008.International Conferences. Denpasar: TEFLIN

O'Malley, J.M and L.V. Pierce. 1996. *Authentic Assessment for English Language Learners: Practical Approaches for Teachers.* London: Longman

Richards, J.C. 2008. *Teaching Listening and Speaking from Theory to Practice.* Cambridge: Cambridge University Press.

Richards, J.C., Hull, J. and Proctor, S. 1997. T*he New Interchange Book 2.* Cambridge: Cambridge University Press.

Sawyer, W., Watson, K. and Adams, A (eds). 1989. *English Teaching from A-Z.* Philadelphia: Open University Press.

Singarimbun, M. and Effendi, S. (ed). 1989. Metode Penelitian Survai. (Revised Ed.). Jakarta: LP3ES.

Tuckman, B.W. 1978. *Conducting Educational Research* (2nd Ed.). New York:Harcourt Brace Jovanovich Publishers.

Weir, C.J.1990. *Communicative Language Testing.* New York: Prentice Hall.