Chabahar Maritime University
Iranian Journal of English for Academic Purposes

IJEAP, (2015) vol. 1 No.4, 26-36 (Previously Published under the title: Maritime English Journal)

English and Non English major Teachers' Assessment of Oral Proficiency: a case of Iranian Maritime English Learners

[1] Hooshang Khoshsima

Associate Professor of TEFL

[2] Ali Asghar Roostami Abusaeidi

Professor of English Literature

ID: 1040

Received: 12 May 2014 Accepted: 11 November 2014 Available online: January 2015

Abstract

Speaking assessment is still construed as a complicated, under-researched process from the vantage point of tasks and rater characteristics. The present study aimed at investigating if and how English Major and none English Major teachers differ in their perception of the construct of oral proficiency while assessing learners' L2 oral proficiency. To this end, 38 male and female non-native EFL teachers were asked to rate 10 monologs on a 4-point rating scale and provide concurrent verbal reports. Cronbach's alpha coefficient shows that the inter-rater reliability is relatively high, however; EM teaches are on the whole more reliable while doing the assessment task. On the other hand MANOVA reveals no significant difference in the teachers' holistic rating of the speech samples ($F=1.44$, $\rho \geq 0.05$), and the adopted approach while doing the assessment task in EM versus NEM teachers' modes of assessment.

Keywords: Assessment, Rater, Performance-Based Assessment, Holistic Rating Scale, Oral English Proficiency Construct, Test

Introduction

1.1 Preview

To many language learners, the ability to speak in a foreign language is equal to being able to use a foreign language. Speaking is a productive skill which involves the process of encoding or creating a message. Like the other skills, speaking is an active skill in which speakers use their background linguistic knowledge to create a meaningful message to the deliberate audience (Chastain, 1988). Speaking skills are an important part of the curriculum in language teaching, and this makes them an important object of assessment as well (Luoma, 2004, p.1).

Shifts from conventional paper-and-pencil selective response tests to performance-based assessment of second and foreign language skills, especially in assessment of writing and speaking, where raters are usually required to carry subjective assessment of a person's language ability, over the past decades has accorded pivotal importance to the role of the raters in assessing students' language abilities. While in conventional tests (i.e. multiple choice tests) the obtained score is the implication of interaction between test task and examinee, in performance-based assessment rater facet is added to the assessment process

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

which can be a potential source of error, influencing test score validity and reliability. Variability caused by raters has been found to manifest itself in a number of ways: raters may differ with regard to the overall internal consistency, they may display different bias patterns, or they may rely on diverse scoring criteria or make different interpretation of rating scales.

An increasing number of studies have focused on rater variability in performance-based assessment of L2 ability. Raters have been proved to differ with regard to the severity of their evaluation of examinees' oral proficiency and can produce a broad range of scores. Raters were also found assigning the same score to disparate performances or disparate scores to the same performance (Brown, 1995; Douglas 1994; Merion & Schi, 2000; Orr, 2002). Some researches revealed that even if high degree of agreement exists between the raters, this does not state by any means similar judgment, in other words the same score may mean different things to different raters (Ang-Aw & Meng Goh, 2011; Douglas, 1994; Johnson & Lim, 2009; Orr, 2002; Merion & Schi, 2000). Raters may also differ in their approaches while assessing speaking. To increase rater consistency and provide a more accurate estimate of examinees' scores, some researchers have recommended rater training sessions. Research findings, however, have shown that although training is effective, it does not eliminate rater variability and rater inconsistency exists even after training programs (Hamilton, Reddle & Spratt, 2001; Knoch, 2011; Lumely. 2002; Weigle, 1998). This lends credence to the use of more than one rater to assess each learner's L2 performance. In many assessment contexts, multiple raters' rating of examinees' performance are combined to produce a single score. But such multi-rater assessment of L2 speaking ability does not usually result in highly reliable and valid scores. There are occasions when raters assign completely discrepant scores to the same performance which requires the use of some method of resolving those differences (Penny & Johnson, 2011).

1.2 Statement of Problem

The expansion in scientific, technical, and economic activity on an international scale after the World War the Second created an appeal for an international language. For many reasons, most significantly the economic power of the United States after World War II, this role fell to English (Hutchinson & Waters, 1987). With the growing demand of learning English as a foreign language over the past decade in Iran, we witness an increasing number of English institutes which require English teachers more than at any other time. The majority of these institutes are private, indicating that they are governed in accordance with their managers' policies. Some institutes just use teachers with English related majors i.e. having a university degree in one of English majors is essential for teaching there. But for some institutes, it is not the case, it does not matter what the teachers' majors are, if their English knowledge is acceptable for teaching English they can take the role of teacher. Parallel with changes occurring in language teaching methods, most of the language institutes focus on communicative ability of the learners, hence, speaking receives vital importance. Since assessment is a part of any teaching curriculum, assessing speaking is of crucial importance especially when learners are going to take part in a placement test the aim of which is to place test takers at an appropriate level in a program or course (Richards & Schmidt, 2002.p.404).

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

As noted earlier, the involvement of raters is a source of error influencing the scores obtained by learners, and in language institutes where the selection of teachers is somehow done without careful consideration; this involvement may lead to even more severe consequences.

1.3 Research Questions

Q1. Is there a significant difference between inter-rater reliability indices in English Major versus None English Major teachers?

Q2. Is there a significant difference between the holistic ratings of 10 speaking tasks by English Major versus None English Major teachers?

Literature review

2.1 Rater Variables

A number of studies have focused on those raters' characteristics that may introduce a source of substantial degree of variability in English as a second language (ESL) and English as a foreign language (EFL) performance-based assessment context. Of these properties, diverse linguistic background and professional experience have received the most attention.

2.1.1 Linguistic Background

As regards raters' linguistic background which has mostly appeared in contrastive studies of native speaker (NS) and non-native speaker (NNS) raters, findings have revealed that although in most cases there are no significance differences between these two groups of raters in the scores they assign to candidates' L2 performance, they differ in their perception of the construct under question i.e. writing or speaking (Fayer & Krasinski, 1987; Santoes, 1988; Shi, 2001; Zhang & Elder, 2011). Zhang and Elder (2011), for instance, found no significant differences between the scores assigned by NS and NNS raters to oral performance of a group of examinees, however, the two groups were found to differ in the way they weighed various aspects of oral proficiency construct. As far as linguistic features were concerned, NNS English teachers demonstrated to be more severe than NS teachers. On the other hand, NS teachers attended more to communication strategies, demeanor and interaction indicating that they based their judgment on how well candidates can accomplish a communicative task rather than on candidates' linguistic competence. It should be noted that drawing clear-cut conclusions about the effect of linguistic background of raters on their rating behavior from extant literature is not safe. Some studies have been carried out with contradictory results. There has been some research, for example, showing that NS raters are likely to be more severe than NNS raters with regard to linguistic features (Barnwell, 1989; Brown, 1995). Another group of studies show no difference between NS and NNS groups with respect to both severity (Johnson& Lim, 2009; Kim, 2009) and consistency (Kim, 2009). These differences can be attributed to different methodologies employed in these studies, small sample size, and diverse native language (Chaulhoub-Deville, 1995; Brown, 1995).

2.1.2 Rating Experience: Novices vs. Experts

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

Rating experience that raters bring to the rating process is another variable that has been found to impact on raters' performance. Research on the effect of rater experience on ESL assessment shows that experienced and novice raters approach the rating task differently (Barkaoui, 2010; Cumming, 1990; Schoonen, Verger & Eiting, 1997; Weigle, 1998). Cumming (1990), for example, found that expert teachers had a fuller mental representation of the task of assessing their students' ESL compositions and used various criteria, self control-strategies and knowledge sources to read and rate the compositions. Novice teachers, on the contrary, employed a few criteria to assess students' compositions and used skills that might derive from their general reading abilities and other sources of knowledge that they had acquired previously such as editing. In a recent study, Barakaoui (2010) also argued that novice and expert raters make differential use of the rating scale. Generally, novices relied more heavily on the rating scale for rating criteria and decision-making because they did not have established criteria and/or they did not know how to approach the rating task. Experienced raters, by contrast, referred to criteria other than those in the rating scale more frequently, gave more comments on the compositions, and were more self-consistent. Taken together, literature on rater experience in ESL assessment shows that raters' expertise is fundamental to their decision-making and does affect their rating performance.

2.1.3 Teaching Experience: Teachers vs. Non-teachers

Some research have made a comparison between assessments done by ESL teachers and those raters without teaching experience (Barnwell, 1989; Brown, 1995; Chaulhoub-Deville, 1995; Hadden, 1991; Shoamy et al 1992). These studies, however, have not provided consistent results as to how teacher and non-teachers' judgments differ. Hadden (1991) figured out that teacher raters were more severe than non-teachers with respect to linguistic ability in assessing speaking ability of Chinese students, but the two groups did not differ significantly as far as comprehensibility, social acceptability and body language were concerned. These findings, however, contrasted with Chaulhoub-Deville's (1995) study which found that non-teachers tended more on linguistic features in a narration task than teachers who referred to creativity and adequacy of information more than linguistic aspects. Chaulhoub (1995) attributed the discrepant findings of the two studies to the different native languages of the participant raters. While in her study, raters were NSs of modern standard of Arabic (MSA), in Hadden's (1991) study the participants were native English speakers.

The above studies address rater variability in performance-based assessment in diverse rater groups. Another line of research, however, has been carried out with the aim of identifying bias patterns among raters, thereby providing a fuller picture of the rater facet in performance-based assessment.

2.2 Rater Training

From research on rater facet in L2 performance-based assessment, it can be understood that there is the possibility of a substantial degree of rater variability in assessing L2 writing and speaking, that raters, consciously and unconsciously, may assess students L2 abilities with bias, and that raters interpret rating scales differently and draw on a range of non-criterion

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

factors which they suppose to be important in assessing oral or written performance(Brown, 2000; Douglas, 1994; May, 2006; Orr, 2002; Wiggleworth, 1993). These among others are reasons that have pushed researchers and administrators towards planning training programs so as to reduce systematic errors caused by raters and improve rater consistency and score validity (Lumely et al., 1995). As such, several studies to date have been carried out investigating whether training can improve raters` performance (Elder et al., 2007; Shohamy, et al., 1992; Tajeddin & Pashmforoosh, 2011; Weigle, 1998). Shohamy et al. (1992) investigated the effectiveness of rater training by comparing rating performance of 10 trained and 10 untrained raters of 50 L2 compositions on three scales: holistic, communicative, and accuracy . They observed that inter-rater reliability was relatively high in both groups, but trained raters were on the whole more reliable than untrained raters.

While the existing literature shows that, on the whole, training reduces rater variability and improves rater self-consistency, it does not appear to eliminate random errors caused by raters, that is, rater inconsistency still exists after regular training session (Lumely et al., 1995, Lumely, 2002; Ang-Aw et al., 2011; Weigle, 1998; Weigle, 1994), receiving individualized feedback(Knoch, 2011; Wiggleworth, 1994) or online self-training programs (Hamilton, Reddle, & Spratt, 2001). In a recent study, Ang-Aw et al. (2011), for instance, investigated rater variability among 7 experienced examiners of 'O' level examination (a high-stakes national English test for secondary students in Singapore (formerly also in the UK, replaced by GCSE). They found that despite undergoing similar training, raters differed in their perception of oral proficiency construct, the emphasis they placed on different aspects of oral proficiency, and their interpretation and approach to assessment. In another study, Knoch (2011) examined the impact of individualized feedback on rating behavior of 19 raters assessing writing and speaking subsets of occupational English Test (OET) over eight administrations. After each administration, raters received a performance profile on their rating behavior on the basis of MFRM. The findings showed that raters rated neither the writing nor the speaking subsets no better after receiving individually targeted feedbacks.

The overall impression gleaned from literature on rater facet in L2 assessment, thus, shows that while rater training which, as stated by Lane and Stone (2006), typically involves "familiarization activities, practice rating, and feedback and discussion" (cited in Lim, 2011, p. 544) can attenuate rater variability, improve self-consistency of individual raters, and reduce rater bias in relation to various aspects of test situations, it does seem to have a temporal effect, usually no more than a day (Congdon & Mcqueen, 2000; Lumely & McNamara, 1995, Weigle, 1998), and does not eliminate the extent of rater variability.For this reason some researchers are against the practice of training raters and conducting judgments on the basis of a single rating by such trained raters and have advocated the use of double or multiple raters, specially in high stakes tests(Lumely et al.,1995). Several studies to date have investigated inter-rater reliability and scoring validity of multi-rater judgments of students L2 performance (Douglas, 1994; Gamaroff, 2000; Meiron et al., 2000), most of them showing that even in cases where high-inter rater reliability is achieved , quantitatively similar scores usually reflect qualitatively different learner performances.

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

Methodology

3.1 Participants

38 Iranian EFL teachers (15 male, 23 female) participated in the study, of whom 19 were EM and 19 NEM teachers. Of the 19 EM teachers, 9 males and 10 females and of 19 NEM teachers, 6 males and 13 females were currently teaching English as a foreign language (EFL) at Safir Language Academy and had been teaching English for a minimum of one year and maximum of 15 years. None of the participants had received any rater training programs. They self rated their English proficiency as either *advanced* or *near native*. Table 1 describes a profile of participants' demographics.

3.2 Instruments

Audio-recording of 10 Iranian EFL learners' oral English were used as the motivated material. They were an unsystematic subsample of a pool of recordings performed for research purposes in an earlier study and were based on 5 topic-based one-way speaking tasks, with 2 students speaking 2-3 minutes on each(Tajeddin, Pashmfroosh, 2011). As noted by (Tajeddin, Pashmfroosh, 2011), the students were adult EFL learners studying English in private institutes.

Participants were also asked to fill an assessment sheet in which they rated speaking samples holistically on a 4-points scale from 1(novice) to 4 (superior) with the midpoints labeled as intermediate and advanced respectively (half points were allowed). As mentioned in the introduction of this chapter no further explanation of the scores was provided to find out how the raters perceived second language oral proficiency construct and defined the scoring criteria (Orr, 2002; Zhang & Elder, 2011; Kim, 2009; Ang-Aw& Meng Goh, 2011). Teachers employed all the scale points, with 4 being the least and 2 being the most used scores in both groups.

3.3 Procedure

38 participants of both genders were selected randomly. The participants were supposed to assess 10 speaking tasks chosen from a pool of recordings of an earlier research (Tajeddin, Pashmfroosh, 2011). The assessment took a holistic rating on a 4-point scale from 1 (novice) to 4 (superior) with the midpoints labeled *intermediate* and *advanced* respectively. Each rater was briefed on the rating scale and speaking samples and received instruction on how to produce think-aloud protocols while rating the recordings. Nothing was said about student's name, specific age, and level of proficiency, but they were told that the speakers were EFL students who spoke on a specified topic after giving one minute to think about it. Teachers, then, rated the 10 recordings on the basis of the holistic 4-point rating scale while thinking aloud into a tape recorder. To eliminate researcher effect on their performance and to provide them with sufficient time, raters were allowed to do the ratings at their convenience. The researcher came up with 38 assessment sheets and 380 sets of verbal protocols.

Data Analysis

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

4.1 Analysis of the First Research Question

The first research question focused on investigating inter-rater reliability indices in EM versus NEM teachers' modes. As noted earlier the 38 participants in the study were asked to complete an assessment sheet in which they rated each speaking sample holistically on a 4-point rating scale, labeled as 1= novice, 2= intermediate, 3= advanced, 4= superior (half points were allowed). Consistency of ratings in both modes - EM teachers vs. NEM teachers was estimated by means of Cronbach's alpha coefficient.

Table 4.1 Inter-rater reliability estimates for EM vs. NEM teachers modes.

Modes of assessment	α coefficient	95% confidence interval	
		Lower Bound	Upper Bound
EM teachers	0.81	0.65	0.91
NEM teachers	0.79.4	0.62	0.90

Inter-rater reliability in the EM teachers' mode and in NEM teachers' mode were 0.81 and 0.79.4 respectively (table 4.1) which is surprising given the fact that the participant teachers had not received any training. The high level of inter-rater consistency in both modes can be attributed to the fact that raters were allowed to augment integer-level scores by using half points. Thus permitting such flexibility in the scores assigned by raters can be suggested as a reason for the high level of consistency between raters in EM and NEM teachers. Although the indices of agreement between raters for each mode are very similar, it is important to look at the confidence intervals for the reliability indices in each mode (table 4.1).The confidence intervals show the range within which the population indices may fall in 95% of samples. For EM teachers' mode, the confidence intervals are 0.65 to 0.91 and for NEM teachers' mode 0.62 to 0.90. These results show narrower confidence intervals for alpha coefficient in the EM teachers' mode of assessment. Although the inter-rater reliability was relatively high in both modes, EM teachers were on the whole more reliable while doing the assessment task.

4.2 Analysis of the Second Research Question

The second research question aimed at investigating the difference between the scores assigned to the 10 speaking tasks by teachers in the two modes of assessment. To achieve this aim SPSS version 17 was used. Table 4.2 summarizes means and standard deviations for teachers` assessment of the 10 speaking samples in the two modes.

Table 4.2 Descriptive statistics for the 10 speaking samples in EM vs. NEM

	NEM mode		**EM mode**		Ranking of samples	
	Mean	S.D.	Mean	S.D.	NEM	EM

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

Task 1.1	2.28	.50	1.83	.38	8	9
Task 1.2	2.60	.60	2.27	.46	3	4
Task 2.1	2.35	.76	2.01	.36	6	7
Task 2.2	3.36	.57	2.82	.68	1	2
Task 3.1	1.93	.73	1.77	.64	10	10
Task 3.2	2.18	.71	2.05	.53	9	6
Task 4.1	2.48	.72	2.16	.38	4	5
Task 4.2	3.31	.60	3.05	.63	2	1
Task 5.1	2.33	.67	1.94	.53	7	8
Task 5.2	2.44	.65	2.50	.70	5	3
Total	2.58	.65	2.24	.50		

As it is shown in table 4.2 in the EM teachers' mode, mean scores for the 10 tasks ranged from 1.77 to 3.05 and in the NEM teachers' mode it ranged from 1.93 to 3.36. The total mean score of the 10 speaking sample for EM teachers' mode is 2.24 and for NEM teachers' mode is 2.58 which shows teachers assigned slightly higher scores to the 10 tasks in NEM mode while doing assessment. Moreover, considering standard deviations, a smaller total mean of standard deviation was found for EM mode (total mean of S.D. of .50 vs. .65). This suggests that scores awarded by EM teachers to the 10 speaking samples were more homogeneous than those they gave in the NEM mode. With respect to the rankings, in the NEM mode, Task 2.2 received the highest score and Task 4.2 the second highest score. In the EM mode, teachers scored Task 4.2 as the best and Task 2.2 as the second best. They agreed in both modes, on the poorest performance by assigning the lowest score to Task 3.1. Concerning the rest of the tasks, a difference of one to two ranks emerged between raters` performance in the two modes.

To find out the differences in the scores assigned by the teachers to 10 speaking tasks in EM and NEM mode, MANOVA was run. Table 4.3 shows the results of MANOVA comparing the scores in EM and NEM mode.

Table 4.3 Results of MANOVA comparing the scores in EM and NEM

Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Wilks` Lambda	.74	1.44	10.000	43.000	.192	.252

The F- observed value is 1.44 and significance level is .192 which is more than 0.05 which shows no significant difference between the scores assigned by teachers to the 10 speaking tasks in EM versus NEM mode(F=1.44, $p \geq 0.05$).

5. Discussion and Conclusion

As indicated in preceding parts and chapters, this study set out to ascertain whether there is any difference between EM and NEM teachers' assessment of L2 oral proficiency of Iranian EFL learners. Four research questions were addressed in this descriptive research that would

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

be discussed here. This study examined teacher variability through both qualitative and quantitative approaches, focusing on inter-rater reliability indices by EM and NEM teachers, the differences between the scores assigned to 10 speaking tasks by them, their perception of oral proficiency construct, and approaches used by them while assessing L2 oral proficiency.

To answer the first research question Cronbach's alpha coefficient was run. As it was shown in table 4.1 inter-rater reliability in the EM teachers' mode and in NEM teachers' mode were 0.81 and 0.79.4 respectively. . Although the indices of agreement between raters for each mode are very similar, it is important to look at the confidence intervals for the reliability indices in each mode (table 4.1). For EM teachers' mode, the confidence intervals are 0.65 to 0.91 and for NEM teachers' mode 0.62 to 0.90. These results show narrower confidence intervals for alpha coefficient in the EM teachers' mode of assessment. Thus, although inter-rater reliability was relatively high in both modes, EM teaches were on the whole more reliable while doing the assessment task. So the null hypothesis assumed for this research question is rejected.

The second research issues under-question in this study was whether EM and NEM teachers differed in their holistic rating while doing the assessment task. Based on the results stated in table 4.2 teachers assigned slightly higher scores to the 10 tasks in NEM mode while doing assessment. Moreover, considering standard deviations, a smaller total mean of standard deviation was found for EM mode (total mean of S.D. of .50 vs. .65). This suggests that scores awarded by EM teachers to the 10 speaking samples were more homogeneous than those they gave in the NEM mode. To discover the differences in the scores assigned by the teachers to 10 speaking tasks in EM and NEM mode, MANOVA was run. according to table 4.3 the F- observed value is 1.44 and significance level is .192 which shows no significant difference between the scores assigned by teachers to the 10 speaking tasks in EM versus NEM mode(F=1.44, $p \geq 0.05$). Therefore, the null hypothesis supposed for this research question is accepted.

5.1 Pedagogical Implication

The overarching aim of the present study was to explore the differences between a group of EM and NEM teachers in assessing candidates' L2 oral proficiency. The findings of the study have a number of implications for teacher educators, and teachers and language institutes.

A crucial implication of this study would be a change in the policies of private language institute in choosing English teachers. The owners of these institutes might be more cautious about recruiting teachers to teach English in their institutes. Another implication concerns with educating programs with the aim of making teachers more homogonous in the assessment of subjective tasks i.e. speaking and writing, in any teaching curriculum including private language institute.

References

Ang-Aw, H., & Meng Goh, C. (2011). Understanding Discrepancies in Rater Judgment on National-Level Oral Examination Tasks. *RECL Journall* .

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

Barkaoui, K. (2010). Variability in ESL Essay Writing Ratin Processes:The Role of the Rating Scale and Rating Experience. *Language Assessment Quarterly*, 54-74.

Branwell, D. (1989). Native Speakers and Judgments of Oral Proficiency in Spanish. *Language Testing*, 152-163.

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation-Specific Language Performance Test. *Language Testing*, 1-15.

Brown, A. (2000). An Investigatigation of the Rating Process in the IELTS Oral Interview. *IELTS Reasearch Report*, 49-84.

Brown, H. D. (2000). *Principles of Language Learning and Teaching*. New York: Pearson Education.

Chastain, K. (1988). *Developing Second Language Skills Theory and Practice*. Florida: Harcourt Brace Jovanovich.

Chaulhoub-Deville, M. (1995). Deriving Oral Assessment Scales Across Different Tests and Rater Groups. *Language Testing*, 16-33.

Elder, G., Barkhuizen, G., Knoc, U., & Randow, J. V. (2007). Evaluating Rater Responses to an Online Training Program for L2 Writing Assessment. *Language Testing, 24*(1), 37-64.

Fayer, J. M., & Krasinki, E. (1987). Native and Non Native Judgments of Intelligibility Andirritation. *Language Learning*, 313-326.

Hadden, B. L. (1991). Teacher and Non-Teacher Perceptions of Second-Language Communication. *Language Learning*, 1-24.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' Perceptions of Online Rater Training and Monitoring. *System, 29*, 505-520.

Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes: A Learning-Centered Approach*. Cambridge: Cambridge University Press.

Johnson, J., & Lim, G. (2009). The Influence of Rater Language Background on Writing Performance Assessment. *Language Testing*, 485-505.

Kim, Y. (2009). An Investigation into Native and Non-Native Teachers' Judgments of Oral English Performance: A Mixed Methods Approach. *Language Testing*, 187-217.

Knoch, U. (2011). Investigating the Effectiveness of Individualized Feedback to Rating Behavior a Longitudinal. *Language Testing, 28*, 2-5.

Lumley, T. (2002). Assessment Criteria in a Large-Scale Writing Test: What Do They Really Mean to the Raters? *Language Testing, 19*(3), 246-276.

Lumley, T., & Mcnamara, T. (1995). Rater Charecteristics and Rater Bias: Implications for Training. *Language Testing*, 54-71

Orr, M. (2005). The FCE Speaking Test: Using Rater Reports to Help Interpret Test Scores. *System*, 143-154.

Penny, A. J., & Johnson, L. R. (2011). The Accuracy of Performance Task Scores After Resolution of Rater Disagreement: A Monte Carlo Study. *Assessing Writing, 16*, 221-236.

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com

Penny, A. J., Johnson, L. R., & Gordon, B. (2000). The Effect of Rating Augmentation on Inter-Rater Reliability: An Emprical Study of an Holistic Rubric. *Assessing Writing, 7*, 143-164.

Richards, J. &., & Schimdt, R. (2002). *Longman Dictionary of Language Teaching and Applied Linguistics (3rd ed.)*. New York: Longman Publications.

Shohamy, E., & Gordon, C. K. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, 27-33.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, 27-33.

Tajeddin, Z., & Pashmforoosh, R. (2011). Non-Native Teachers` Criteria for Speaking: Does Rater Training Make a Difference? *The First TESOL Persia Conference*, (pp. 1-10).

Weigle, S. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, 263-287.

weigle, S. C. (1994). Effects of Training on Raters of ESL Compositions. *Language Testing, 11*, 197-223.

Zhang, Y. & Elder (2011). Judgment of Oral Proficiency by Non-Native and Native English Speaking Teacher Raters: Competeting or Complementary Constructs? *Language Testing*, 31-50.

[1] Corresponding Author;
English Language Department, Chabahar Maritime University, IRAN Email: khoshsima@cmu.ac.ir
[2] Ministry of Science & Technology, IRAN Email: rostamiabu110@yahoo.com