Advances in Language and Literary Studies
ISSN: 2203-4714
www.alls.aiac.org.au

Assessing Individual and Group Oral Exams: Scoring Criteria and Rater Interaction

Özlem Yalçın-Çolakoğlu[1]*, Merve Selçuk[2]

[1]*School of Foreign Languages, Bahçeşehir University Çırağan road, Osmanpaşa Mektebi Street, 34353 Beşiktaş-İstanbul, Turkey*
[2]*School of Foreign Languages, Altınbaş University Mahmutbey Dilmenler Street, Mahmutbey Campus 26, 34218 Bağcılar-İstanbul, Turkey*

Corresponding Author: Özlem Yalçın-Çolakoğlu, E-mail: ozlem.yalcin@sfl.bau.edu.tr

ARTICLE INFO

Article history
Received: October 06, 2018
Accepted: January 17, 2019
Published: February 28, 2019
Volume: 10 Issue: 1
Advance access: January 2019

Conflicts of interest: None
Funding: None

Key words:
Individual and Group Tasks,
Rater,
Rubric,
Scoring criteria,
Speaking

ABSTRACT

Criterion referenced tests of second language speaking performance are administered in different institutions using different procedures. The present study reports raters' practices of second language speaking tests, in particular the correspondence between test-takers' grades when assessed individually and in groups. Data derived from audio-recordings of raters' (n=8) decision making process (scoring) in two test modes, post-test interviews and two sets of students' (n=92) speaking scores were obtained from individual versus group discussion tasks. Although a grading rubric had been used, it was found that raters also relied on rubric irrelevant criteria when judging performances, which brings up the question whether the validity of the inferences is jeopardized.

INTRODUCTION

Today, the field of language testing has moved from its traditional roots of factual knowledge testing towards "assessment for learning" with a strong interest on performance assessment. According to Jonsson & Svingby (2007) "performance assessments are designed to capture more elusive aspects of learning by letting the students solve realistic or authentic problems." However, the assessment of performance, when compared to e.g. multiple-choice assessment, reveals the problem of trustworthy decision-making processes. In this sense, the assessment of speaking proficiency in the second or foreign language teaching context requires ratings by trained raters. Scores on such examinations may vary due to differences in speaker performances as well as rater performances. As the interest on performance measures is increasing, the issue of reliability and validity still remains to be a highly debated topic (Van Moere, 2012). In the field of language assessment, variation in rater judgments is sometimes regarded as a source of bias (Holzbach, 1978). This issue, also known as score variability, dates back to Lado (1961), who compared selected-response versus constructed-response tasks and highlighted the significance of rater reliability by saying "In objective tests, scorer fluctuation is practically nil and need not be considered a factor. In various production tests such as essays or long response tests, scorer fluctuation can be a major factor of the unreliability of a test" (p. 331). In line with Lado, Bachman et al. (1995) argued that even though performance language tasks are more representative of real-life language use, we cannot ignore the fact that rater judgments and variability in tasks are a major source of measurement error and for this reason, threaten the reliability and validity of test scores. Therefore, raters are regarded as central as performers in productive language assessment (McNamara, 1997).

LITERATURE REVIEW

Reliability in Scoring

Most exams have consequences for all stakeholders involved in the assessment process (Black, 1998). Therefore, the evaluation or scoring of a performance has to be trustworthy, relying on disinterested judgment and based on some kind of evidence (Wiggins, 1998). To reach a fair judgment, the assessment should be free of bias, meaning independent of the rater and regardless of the time and place of assessment. However, according to Jonsson & Svingby (2007), this is hardly obtainable. Whereas this is not a problem in traditional testing, e.g. multiple-choice type exams, it is questioned

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.alls.v.10n.1p.147

a lot in performance assessment on behalf of its credibility, focusing mainly on the reliability of rater's measurement. It is said that the more consistent the scores are, when assessed by different raters and in different occasions, the more reliable the assessment is believed to be (Moskal & Leydens, 2000). Variations in rater judgments can be discussed in terms of inter-rater reliability, consistency across different raters, and intra-rater reliability, the consistency of test scores given by one single rater. According to Lumley & McNamara (1995), errors in measurement can be directly related to rater biases which can be observed in raters' behaviors, such as raters displaying "particular patterns of harshness or leniency in relation to only one group of candidates, not others, or in relation to particular tasks, not others, or on one rating occasion, not the next" (p.56).

For the purposes of the study, variation in rater judgment is mainly analyzed through raters' obedience to an evaluation criterion known as a scoring rubric and through the variance in speaking tasks, namely individual vs. group.

Scoring Rubrics

To minimize the possibility of rater bias, most EFL/ESL speaking exams are carried out with the help of assessment tools known as scoring rubrics to guide raters in making their judgments based on certain standard criteria. These criteria are established with an underlying theoretical framework that defines the construct and outlines the components of speaking ability (e.g., fluency, vocabulary, cohesion, grammar, etc.). Rubrics play an important role for both sides, assessor and examinee, of the assessment process as they provide a general idea of what assessors should listen for (Arter & McTighe, 2001; Busching, 1998; Perlman, 2003). Moreover, the main reason for using rubrics is: a) to provide evidence that ratings are related to the theoretical underpinnings, and b) to ensure inter-rater reliability (Douglas, 1994). Usually, raters are provided with training sessions in which they are introduced to the scoring rubric and asked to assess some sample performances. The raters are first expected to assess individually and then to compare with other raters to see how consistent or, vice versa, inconsistent they are. Through these trainings, raters' awareness about the salient features of the scoring criteria is raised. The use of rubrics provides transparency for the assessment, meaning that the expectations are made explicit through well-defined criteria. Not only does it enable transparency, but it also facilitates the process of providing feedback and helps for self-assessment.

Rubrics can be divided into two categories: analytic and holistic. In analytic rubrics, raters rate the performance of the examinee based on each category of the rubric whereas in holistic rubrics, raters rate the performance on an overall judgment. While analytic scoring is more preferred in classroom assessment due to the fact of its diagnostic nature, holistic scoring on the other hand, is generally used for large-scale assessment, as it is easier and less time-consuming. In any case, regardless of the nature of the scoring rubric, raters are expected to apply the criteria consistently across all subjects whose performances are tested. More than that, one of the most important reasons for using a scoring rubric is to prevent raters from comparing an examinee's performance with another examinee's and to solely make them refer to the scoring rubric while making a decision (Orr, 2002).

Individual vs. Group Speaking Tasks

Most of the performance-based speaking tasks take place in an individual format, meaning a traditional face-to-face interview between an interlocutor and an examinee. In addition to this task type, many institutions, since the late 1980s (Ducasse & Brown, 2009), prefer to administer pair and group discussion tasks in which examinees are tested together with their peers, but assessed on their individual performance. According to Kramsch (1986), interaction needs to be regarded as a major component of speaking ability revealing speakers' competence in speech sequencing and implementation of turn-taking rules. In addition to that, Van Lier (1989) states that the main reason for preferring group tasks has been the idea to move away from interview type tasks in which students were assessed on "test discourse" and "institutional talk" rather than normal conversation and interaction. This shift from individual to peer assessment has revealed positive washback on the classroom (Messick, 1996). Moreover, this form of assessment has been reported as appropriate for some situations or as one part of a task battery (Fulcher, 1996; Shohamy et al, 1986) and discussed to be more of use for large-scale testing, such as in schools and universities, as three to six examinees can be tested simultaneously, making it more economical (Bonk & Ockey, 2003; Folland & Robertson, 1976). Even though this pair or group assessment is regarded as an effective task for assessing interactional skills of an examinee, there has been little research about raters' judgments of these group performances (Ducasse & Brown, 2009). In fact, understanding what raters value while judging examinees' performances is one of the most important aspects of group assessments, as it is a rater's view of interaction that is embodied in the test score. Therefore, it can be claimed that test validity mostly depends on the rating process and the evaluated criteria, as well as on the task performance. As claimed by Brown, "...in any assessment involving judgment, it is the criteria by which the performance is judged which define the construct" (2005, p.26). However, most of the peer and group task studies have mainly focused on the relationship between test scores and testees' characteristics (e.g. O'Sullivan, 2002; Norton, 2005). One remarkable study that needs to be mentioned though is that of May (2006 a, 2006 b), in which retrospective verbal reports were used to investigate raters' judgments of paired discussion tasks. The result of the study indicated that raters considered examinees' body language, assertiveness during communication, and their ability to handle the discussion and work cooperatively as part of their judgment.

Purpose of the Study

One of the main issues in language assessment is validity and reliability of test scores. Validity in oral examination is concerned with whether test scores serve the purposes they are intended to. To maintain reliability, raters need to be consistent with rubrics while assessing spoken performance (Luoma, 2004). Therefore, the main purpose was to uncloak raters' "implicit criteria" (Brown, 2000) when assessing students' performance during individual and group tasks in their second language (L2). For this, the study addressed the following research questions:

1. Which factors are involved in the decision-making process of raters when assessing individual and group discussion tasks?
2. Is there a significant difference in students' test scores in terms of individual performance and group discussion performance?
3. What are the perceptions of raters regarding the two assessment procedures?

METHOD

Setting and Participants

The research project took place at a language preparatory school of a foundation university in Istanbul, Turkey. 92 EFL learners with L1 Turkish background at B2 level and 8 raters participated in this study. The raters who may be characterized as a 'convenience sample' consisted of EFL instructors with at least 5 years of teaching experience. They assessed students in pairs, each consisting of one native and one non-native EFL instructor. So, in total 4 pairs of raters assessed 92 students in the first session individually and then the same examinees in groups of three (Table 1). The raters were not informed about the purpose of the project so as not to jeopardize the validity of the study.

Instrumentation

Various data collection tools were used to triangulate the data. Quantitative data came from students' test scores of individual and group task speaking performances. Qualitative data, on the other hand, were collected through verbal protocols, a preferred tool of collecting data in speaking tests to examine the decision-making process of raters more closely (Ducasse & Brown; 2009; Ericsson & Simon, 1993; Green, 1998). They included recordings of examinees' responses to individual and group prompts and rater discussions during scoring to understand thought processes. In addition to students' test scores and recordings of the decision-making process, the study also relies on the perceptions of raters. Data on raters' perceptions were collected using semi-structured interviews, with a set of common questions at the beginning and the end of each interview. The interviews, based on recordings of raters' decision-making process, were thought to present rich data to identify the immediate follow-up of issues emerging in the recordings.

Table 1. Number of participants

Raters	Number of test takers	
	Individual task	Group task (group/ss)
Pair 1 Raters S & Sr	20	7/20
Pair 2 Raters C & Me	22	7/22
Pair 3 Raters P & E	24	8/24
Pair 4 Raters U & M	26	9/26

Procedures

The oral assessment component consisted of two tests: individual and group oral tasks. The former required the examinee to speak on a given prompt for 2 minutes. In group speaking assessment, on the other hand, three or four examinees were awarded scores on their ability to discuss a given prompt, a format for assessing the speaking ability of EFL test takers.

Individual speaking task structure

The student picks a card with a topic followed by three questions. S/he is given 30 seconds thinking time and encouraged to speak about the topic for 2 minutes.

Group speaking task structure

One of the students picks a card with a topic followed by three questions. S/he reads it aloud to the group members and they are given 30 seconds thinking time. All are encouraged to discuss about the topic for 6 minutes.

Interviews

To address the perceptions of raters regarding the assessment procedures, semi-structured interviews were conducted by a single researcher. Prior to the interviews, the researchers transcribed the verbal protocols of raters' negotiations during the scoring process to investigate raters' adherence to the scoring rubric. Additionally, an open coding method was used to identify raters' judgments deemed to be irrelevant to the rubric (e.g. topic difficulty) on student performances. The identified codes were used as questions during the rater interviews to reveal the raters' perceptions pertaining to the two different speaking tasks. The interviews were audio-recorded and each lasted between 7 to 10 minutes.

RESULTS AND DISCUSSION

Analysis

In an effort to measure the decision-making process of raters in assessing individual and group tasks, approximately 16 hours of audio recorded data were listened to, but only raters' discussions were transcribed. Due to the exploratory nature of the study, a bottom-up approach to coding was adopted. The transcribed data were analyzed using an inductive approach in which themes and patterns emerged from the data. First, codes were identified individually by each researcher and then they collaborated to come up with themes for these codes. The codes are based on comments including the reasons for the scores that the raters assigned according to the

Table 2. Summary of coded data from the scoring process

Coding category	No of raters		No of incidents	
	Individual task	Group task	Individual task	Group task
Reference to rating criteria	8	8	92	92

Table 3. Raters' comments on students' performance

Rater	Comments
E	"Arzu was a bit better in coherence" "İrem and Arzu were more open to interaction." "I think the best was Sumeyye." "Özge was better."
P	"Nilay was more talkative." "Dilber had more mistakes. But she spoke more." "Alper was better than the others."
Ma	"The previous group did better" "They communicated well, not as good as the previous group though" "Zeynep has weaker vocab than Mehmet" "Compared to Saadet, Nur Irem was 1 or 2 points lower" "Aysenur did a little bit better than Aylin" "Aylin was more efficient in terms of fluency"
C	"Irem is a quieter student and she did not interact as well as Saadet Nur" "Zulal excels a bit better in vocab and that is why she scored a bit better than fellow friends" "I will give Aylin considerably lower than the others" "Compared to Mehmet she was lacking those skills"
Me	"I will give Bahar the lowest for vocabulary." "In grammar Melissa was better than the others" "Rümeysa was the most fluent one" "Selen was worse than the others"
U	"Since Edanur was the leader (lead the discussion) she gets the highest score." "I deduct points because she was not as fluent as Gamzenur" "Zeynep's grammar was not ok compared to Mehmet" "In this group Gamzenur shines" "This is by far the worst group we had" "They are better than the previous group" "Mehmet is better than Zeynep" "They were all nearly the same so I will give all the same grade"
Se	"Interaction was the best in this group" "I think though Hilal was the best"
Ser	"They were better than Tahire so I gave them higher score" "Interaction, who was the best?" "Tahire was quieter" "Yes, she was the best"

scoring rubric. Likewise, semi-structured interviews were analyzed following the same procedure.

Quantitative data, on the other hand, included two sets of scores consisting of individual and group performance tasks of the 92 test takers, awarded by 8 raters. The data were analyzed with a paired t-test to examine whether there were significant mean differences based on the students' two test scores. Additionally, correlation analysis was computed to figure out the consistency of scores over time using different scoring procedures.

Results

The first research question aimed at identifying factors involved in the rating practices of raters while judging individual and group discussion task performances. Data analysis indicated that raters used both scoring rubric and other criteria. The following sections present patterns which were identified.

Reference to scoring rubric

As Table 2 illustrates, all raters (n=8) referred to the scoring rubric while making decisions in all cases (N= 92) in both individual and group tasks which shows that they discussed each student individually according to the rubric. In the analysis of rater discussions, it was found that they interpreted students' performance using some other criteria that are non-existent in the rubric such as gestures, eye-contact, and body language. In semi-structured interviews, the raters were asked if they employed different criteria in their mind while assessing individual and group task performance. Some of the comments were as follows.

> ***Rater Se:*** *Well, yeees..., I use different criteria in my mind when I assess students in groups. I did not think about it but yes. I have never thought about it.*
> ***Rater M:*** *I pay attention to gestures, small responses like "aha, yes, well" anything appropriate, body language and anything like that in group tasks.*
> ***Rater P:*** *I consider eye-contact, body language...,*

Reference to rating criteria other than rubric

Analyses display that raters also followed other paths rather than utilizing rubric while assessing the examines. They compared the test takers to each other and rearranged the scores they assigned to each, displayed leniency or severity based of their perceptions of prompt difficulty level or did not comply with time limit. Comparison of test takers, rater leniency or severity and adherence to time limit emerge to be three criteria that raters use other than the scoring rubric.

Comparison of Test Takers

When attempting to identify the salient decision-making process points, a common element was that all rater pairs made a comparison among test takers when scoring the performance (Table 3). They compared mostly individuals in the same group to each other and graded the examinee a higher or lower rating than they deserved according to the scoring rubric.

One pair, however, also compared groups to each other. The set of extracts below provides examples of rater comments.

All in all, all eight raters made comparisons for 68 students out of 92 in the group assessment task. In some cases, they changed the test takers' scores after discussions among themselves. For example, a pair added up the score and realized that "it is the same for everyone" but since they believed one of the students in the group was worse, they decreased her overall score. However, in individual tasks, only two of the raters made comparison among eight students.

In semi-structured interviews, the aim was to figure out whether the raters themselves would comment on off-rubric thinking and whether they were aware that such a bias existed. Through the qualitative analysis it was revealed that most of the raters knew that they compared the test takers' performances. When reminded that they were comparing the test takers' performance while scoring, they stated the following:

> ***Rater Se:*** *I have never thought about that but yes I do compare students, other student is kinda like my bottom line. You know, if this student can use this, they are in the same class, the same level, then this student should be using the same thing as well. (...) Nooo, IT IS NOT FAIR, everyone is different, but, yeah, but I think I do that in my mind. (...) No, not when I assess them individually, though.*
>
> ***Rater C:*** *You kinda like to see the difference in their levels, maybe it is not really fair because it depends on the group that they are in. (...) No I do not compare them when they are on their own.*
>
> ***Raters P and E:*** *Yes we compare students in their own group though we assess them individually as well.*

However, one pair of raters, disagreed with the notion that it was unfair and reported that they did this on purpose for scoring consistency. Raters Ma and U commented on their awareness of comparing test takers as follows:

> ***Rater M:*** *We compared them individually also in terms of the whole group. They are judged against their peers, I think it is easier for us to judge their abilities, because we are doing it anyway in terms of the whole class. (...) It is fair and almost necessary. (...) As well, the rubric also mirrors that, the students that are performing the best are the ones that we ticked off in the rubric in the highest section.*
>
> ***Rater U:*** *We tried to do so when assessing them individually but then we can only remember the best and the lowest ones. In group tasks it is easier.*

Rater leniency or severity

There were incidents where raters had discussed the difficulty level of the questions and rated students accordingly. Three pairs on 12 occasions exercised leniency or severity, depending on the speaking prompt. In some cases, when they felt that the prompt was difficult, they were inclined to give higher grades which means that they were more lenient towards the test takers who they felt picked harder speaking prompts. Four of the raters in individual tasks, two in group tasks awarded scores to twelve and twenty-six students respectively taking prompt difficulty into consideration. However, correlation analysis did not imply that this affected total consistency of test scores.

Adherence to time limit

The analysis of rater audios revealed that some of the raters strictly followed the exam instructions for the time limit whereas others were more flexible. It was observed that the latter exceeded the time limit in group tasks when they felt the need to encourage less productive students to speak more. All respondents justified this behavior by referring to the differences in turns taken by group members. They admitted that they had a hard time arranging fair time allocation to test takers since they tried to give time for more quiet examinees, and this resulted in longer test time. While raters allocated equal amounts of time to each test taker during individual task performance, they could not arrange the time well in group tasks. While all raters adhered to time limit in individual tasks, none followed the same routine in peer-to-peer interaction tasks. While they gave the same amount of talking time to 78 examinees in the first task, this figure turned out to be only 21 oral exam takers in the group task. The raters' comments in semi-structured interviews also supported this observation.

> ***Rater E:*** *We could not arrange the time in group work and all groups lasted longer. Usually one dominated the group and when the time was up, the shy one did not even utter a word so we had to give extra time to be fair.*
>
> ***R Se:*** *When one test taker speaks more, we had to extend discussion time to be fair to the other members of the group.*

The second research question investigated whether there was a significant difference in students' test scores in terms of individual performance and group task performance. The t-test results indicate that group means are different in both tasks. A paired t-test was performed to see discrepancy between EFL test takers' individual and group performance test scores. The paired sample correlation was.526 (p <.001) indicating a positive relationship between test scores. It should also be noted that moderate levels of reliability of the raters over two administrations were achieved, with reliability estimates of r =.502. As can be seen from the means (Table 4), either the raters were more lenient or students were more productive in the group task.

The third research question investigated raters' perceptions regarding two assessment procedures. All in all, it was observed that EFL teachers have a positive opinion about assessing students in groups. The majority – 7 out of 8 raters – reported that students did better in group tasks than they did individually. They believe that in group tasks, high achievers supported their peers which led to an increase in talking time of low achievers. The following excerpt exemplifies this:

> ***Rater M:*** *"I think students actually performed better in a group than they do individually. I believe that they get nervous when they are one-to-one with the teacher, but*

Table 4. Paired samples t-test of individual vs group task

	Mean	N	Standard deviation	Standard error mean
Individual task	10.27	92	2.05	0.21
Group task	11.81	92	1.66	0.17

> *they are relaxed when they are with their comrades and they actually build on each other's answers. It was much more efficient."*
>
> ***Rater U:*** *(...) When they are one to one with the teacher, we see that the student is shaking.*
>
> ***Rater C:*** *They are more relaxed in group task. For example, Hilal, she is terrible but she did much better in group task because the first time (individual task) she could not speak at all, I know her and she mostly goes blank. Maybe they try harder because they do not want to let their friends down.*

Number of students in group tasks

Throughout the analysis of the qualitative data, it was observed that raters experienced difficulty in scoring test takers in groups. For example, the qualitative results revealed that 6 of the 8 raters complained that they had to take notes all the time while scoring group tasks since they did not know the test takers individually, which made rating exhausting for them. They had difficulty in focusing on individuals in group tasks while the discussion was going on. Concerning the ideal number of students for group tasks, data from answers to semi-structured interview questions indicated that the majority of respondents agreed that groups tasks should not be carried out with more than three students. Only one pair indicated that it should be carried out in groups of four to increase the level of interaction among examinees. They justified their arguments with the following statements.

> ***Rater M:*** *Two students is a bit less, one student becomes dominant in that situation and they start talking as if they were the only student until she is finished. It is not as organic and spontaneous as they are in three.*
>
> ***Rater N:*** *What is the use of having more than 2? There is already interaction between two, what is the justification for more?*
>
> ***Rater E:*** *We should have only two because when two interact, a third one stays outside if s/he is especially shy.*
>
> ***Rater Se:*** *With 3, one of them is always silent. I always pushed that person, because he was listening and he was understanding everything. But with two, they had to speak. Still, I cannot decide whether to have 2 or three because three is better for more interaction, like a class discussion, like in their faculty classes.*
>
> ***Rater P:*** *I had hard time assessing three students at the same time because you know you need to consider so many variables. It is mentally tiring.*

CONCLUSION

This research has provided insight into what raters consider as criteria when assessing examinees during individual and group speaking tasks. The results indicated that there were several factors involved in the variation of raters' judgments. Verbal reports of raters' decision-making processes revealed difficulty in adhering to the scoring rubric. Specifically, three major instances resulted in awarding the test taker a higher or lower rating than they deserved according to the scoring rubric. First of all, all raters compared test takers' performances to others in group performance tasks in most of the incidents. Comparison of test takers' performances emerged as a construct-irrelevant variance in the ratings of oral proficiency tests. Another incident that violated raters' adherence to the rubric was their opinions about the speaking prompts. They tended to score harsher or more lenient while scoring according to their opinions of the prompt's level of difficulty. Finally, raters revealed that they were inclined to give more time to silent group members in order to encourage them. These findings are consistent with Michael Orr (2002) who stated that the verbal reports of many raters show difficulty in adhering to the assessment criteria.

Contrary to Bonk & Ockey (2003) and Folland & Robertson (1976), the results of this study indicate that assessing several test takers at the same time may not save time. It was observed that assessing students individually took less time per student than when they were scored in groups. In semi-structured interviews, the informants' comments indicated that they were inclined to give extra time for group discussion to be fair to other test takers since one examinee might have the potential to dominate the group discussion. However, raters should be properly trained not to extend time meaning that they should allocate equal amounts of time to each student regardless of students' performances during the test.

One of the limitations of the study could be related to the sequencing of the task types since there was one month time lapse between individual and group performance tasks. Higher grades in group tasks could be related to "learning" and "improving" of the speaking skill. One control for limitation could be to conduct the study with a reversed order. Second, though the study was carried out with experienced teachers, it would be interesting to look at the differences between experienced and expert teachers to see whether expert teachers are more consistent in their scoring. The final limitation of the research concerns scorer severity. Raters argued that they scored based on the difficulty of prompts and the students' states (shy vs. confident). However, no investigation was carried out on whether they were consistently lenient or severe in these situations.

To conclude, raters need to be trained on understanding what is sufficient and or insufficient when evaluating performances in regard to the evaluation criteria. As a helpful activity, raters might be trained with samples of expert raters, assessing performances and justifying their scores based on their inferences. This might be an effective awareness raising task to make raters understand how they differ in their decision making from experts. The main recommendation coming from this study is that raters should receive training for rating especially group tasks. During training sessions, the raters need to extensively study the rubric and carry out rating processes with previously rated benchmark samples to ensure validity of the scores. The main goal should be to provide training on issues like: a) coping with the task of awarding scores to individuals in group performance, b) taking notes while assessing each individual in groups, c) arranging time, d) avoiding comparison within the group, and e) justifying their rating decisions.

REFERENCES

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign languages speaking. Language Testing, 12(2), 238–257.

Black, P. J. (1998). *Testing, friend or foe?: the theory and practice of assessment and testing*. Psychology Press.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. P. Lang.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), IELTS research reports (vol. 3, pp. 49-84). Canberra, Australia: IELTS Australia.

Busching, B. (1998). Grading inquiry projects. *New directions for teaching and learning*, *1998*(74), 89-96.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(2), 125-144.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction.*Language Testing*, *26*(3), 423-443.

Ericsson, K., & Simon, H. (1993). Protocol analysis: Verbal reports as data (revised edition). Cambridge, MA: MIT Press.

Folland, D., & Robertson, D. (1976). Towards Objectivity in Group Oral Testing. *English Language Teaching Journal*, *30*(2), 156-167.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing*, *13*(1), 23-51.

Green, A. (1998). Verbal Protocol analysis in language testing research: A handbook (Vol. 5). Cambridge: Cambridge University Press.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366-372.

Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book. New York: McGraw Hill.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Luoma, S. (2004). Assessing speaking. Ernst Klett Sprachen.

May, L. (2006a). An examination of of rater orientations on a paired candidate discussion task through stimulated recall. Melbourne Papers in Language Testing, 11(1), 29–51.

May, L. (2006b). 'Effective interaction' in a paired candidate EAP speaking test. Paper presented at the 28th Annual Language Testing Research Colloquium in Melbourne, Australia, July 2006.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series*,*1996*(1), i-18.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation, 7, 71–81.

McNamara, T. F. (1997). 'Interaction in second language performance assessment: Whose performance? *Applied linguistics*, *18*(4), 446-466.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT journal*, *59*(4), 287-297.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*(1), 33-56.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal*, *40*(3), 212-220.

Perlman, C. C. (2003). Performance Assessment: Designing Appropriate Performance Tasks and Scoring Rubrics.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOl Quarterly*, 489-508.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), 325–344.

Wiggins, G. (1998). Educative assessment. San Francisco: Jossey-Bass.