Applied Research on English Language

V. 5 N. 2 2016
pp:191–210
http://uijs.ui.ac.ir/are

Factors Influencing Iranian Untrained EFL Raters Rating Group Oral Discussion Tasks: A Mixed Methods design

Nasimeh Nouhi Jadesi [1]*, Seyyed Ayatollah Razmjoo [2], Alireza Ahmadi [3]

[1] *Ph.D. Candidate, Faculty of Foreign Languages and Linguistics, Shiraz University, Iran*

[2] *Professor, Faculty of Foreign Languages and Linguistics, Shiraz University, Iran*

[3] *Professor, Faculty of Foreign Languages and Linguistics, Shiraz University, Iran*

Received: 09/05/2016 **Accepted: 17/09/2016**

Abstract: Using a mixed methods design, the present study attempted to identify the factors influencing Iranian untrained EFL raters in rating group oral discussion tasks. To fulfil this aim, 16 language learners of varying proficiency levels were selected and randomly assigned to groups of four and performed a group discussion task. Thirty two untrained raters were also selected based on their volunteer participations. They listened to the audio files of the group discussions and assigned a score of one to six to each language learners based on their own judgements. They also provided comments on each language learners' performance pointing to why they assigned such scores. The researchers had an interview with the raters after the rating session as well. The quantitative phase investigated whether linguistic features of accuracy, fluency, complexity and amount of talk were attended to by the raters in terms of having any relationship to the scores the raters assigned. Speech rate as an index of fluency and amount of talk turned out to be significantly correlated with the scores. Of more importance was the qualitative phase with the aim of identifying other factors that may account for the scores. The comments provided by the raters on each score and the interviews were codified based on Content Analysis (CA) approach. It was found that the raters attend not only to the linguistic features in rating oral group discussions, but they are also sensitive to the interactional features like the roles the participants take in groups tasks and the overall interaction patterns of the groups. The findings of this study may shed light on group oral assessment in terms of training the raters rating group oral tests and developing rating scales specific for group oral assessment.

Keywords: group oral assessment; linguistic features; interaction; mixed methods design; content analysis

---

* **Corresponding Author.**

Authors' Email Address:
[1] **Nasimeh Nouhi Jadesi** (nnoohij1985@gmail.com), [2] **Seyyed Ayatollah Razmjoo** (arazmjoo@rose.shirazu.ac.ir), **Alireza Ahmadi** (arahmadi@shirazu.ac.ir)

ISSN (Online): 2322-5343**, ISSN (Print):** 2252-0198

Introduction

The interactional nature of oral language use has led to an increase in the incorporation of group discussion tasks. In the last couple of decades, the direct assessment of spoken language has seen a shift in interest towards tests in which test takers interact with each other, rather than with an interviewer. This shift reflects a transition from conceiving of speaking ability as represented by the linguistic features of an individual's spoken words to one of interactive communication. The initiative to incorporate group tasks into the study is to help reflect the diversity existing in the daily use of language.

Group tasks have an orientation toward the social dimension of interaction in second language oral assessment. Coining the term *interactional competence*, Kramsch (1986) conceptually attributed to Vygotsky's (1978) sociocultural theory. He argued that "successful interaction presupposes … the construction of a shared internal context … that is built through the collaborative effort of the interactional partners" (p. 367).

Group oral test format is favoured due to several advantages it offers as an oral assessment task. The first advantage is that it is relatively practical (Ockey, 2001) since more than one test taker can be assessed at the same time, and also raters do not need specialized training for how to conduct effective interviews. The second is that the group oral offers the potential of positive washback for communicative language teaching purposes (Hilsdon, 1995). Since no intrusion or prompting is made by the rater, another important practical advantage is the fact that test administrations are potentially uniform across raters, hence securing the validity of the test. The group oral discussion task, is designed in a way that it yields authentic discourse, since test takers are expected to have discussions similar to those they might have in the real world.

Review of Literature:

Group oral assessment

The results of the studies carried out on the group oral assessment are contradictory. In fact, some of them revealed that this task type can lead to valid score interpretations and some others showed that it does not. There are several studies that seem to confirm the validity of group oral discussion task.

Bonk and Ockey (2003) concluded that the group oral does have potential for yielding valid score-based inferences. Fulcher (1996) showed that variance contributed by task type was negligible, and since fit statistics on a partial credit Rasch model indicated that all three tasks were operating on a one-dimensional scale, they were presumably tapping the same

language knowledge or skills. Arguing for the validity of oral group discussion task, Van Moere (2006) examined scores produced on a large-scale group oral performance test showed that they are useful for making general inferences about a candidate's ability of oral second language proficiency.

However, there are studies that shed doubt on the validity of the group discussion as a speaking assessment task. Two studies questioned the validity of the score-based inferences yielded from the group oral. He and Dai (2006), indicate that at least in certain contexts, the validity of the score interpretations yielded from the group oral are suspect. In the same vein, Shohamy *et al*. (1986) speculate that the group task elicited a different range of language to one-on-one interviews, and added to their claim that a group test should be included as one part of an oral test battery. The results of these studies may challenge the assertion that the group discussion tends to produce natural and extended conversation, which some maintain is appropriate for the all-round display of speaking ability in context (Van Lier, 1989). However, one point missing in this regard is the issue of how the task is implemented. The researchers should set the design and procedure in a way the test takers do believe in the authenticity of the situation.

Oral assessment and linguistic features of speech samples

Another important line of research in oral assessment is the linguistic features of the speech samples produced. The most important linguistic features of speech sample referred to in literature are accuracy, fluency and complexity which are abbreviated as CAF. They are the most widely used measures of oral proficiency.

Several studies have been carried out which investigate the linguistic features of the speech sample and the scores assigned to them. Although they all investigate how CAF measures predicts the overall speaking proficiency, each tap on different related issue with varying variables, methods and instruments (Iwashita, 2008; Ginther, Dimova, & Yang *et al.*, 2010)

Iwashita (2010) nicely summarizes the studies done on linguistic features and oral proficiency scores:

'A considerable number of studies have investigated features of oral proficiency using various methods. The results differ, however, depending on the data type and the methodology. That is, from studies that use data in the form of ratings and feedback on ratings, grammatical accuracy is the principal determining factor for raters assigning a global score, with some variation in the contribution of other factors depending on proficiency level.

On the other hand, in studies that conduct in-built analyses of learner performance, vocabulary and fluency are the principal factors, but, depending on the level, other features come into play'(p.5).

Oral assessment and extra linguistic features

The two studies stated below, are among the studies which have used Content Analysis to come to a more meaningful picture of group oral assessment task analyses revealing an in-depth understanding about the underrepresented features accounting for the scores the raters assign.

Lazaraton and Davis (2008) examined test takers' discourse features to pinpoint discourse features that could account for the scores assigned by the raters. By providing turn-by-turn interactional codings, the authors showed that paired discussion enabled test takers to position themselves as being *proficient, interactive, supportive,* and *assertive*. The findings showed that "language proficiency identity may be locally constructed, mediated, and displayed by test takers in their task talk" (Lazaraton & Davis, 2008, p. 329). The findings revealed that proficiency is fluid and changing depending on the interlocutor and the identity resources they brings to the interaction, which indicates interlocutor influence on candidates' oral performance (cited in Sun, 2014).

Luk (2010) conducted a comprehensive investigation of interactional features in a group oral assessment. The results revealed eight key discourse features reflecting test takers' attempt to gain a high scores to present themselves as efficient speech partner and not caring about an authentic communication.

As evident in the these studies**,** using micro-analytic approaches like CA can provide an in-depth and fine-grained description of the interactional dynamics available in paired and groups oral tasks.

Purpose of the study:

Despite their several merits, as mentioned above, group oral tasks, as an oral assessment task type, have not received the attention they deserve among researchers in terms of the raters rating such tasks. Being human, raters as an important facet in oral proficiency assessment, are inevitably subject to a wide range of factors that may reinforce or threaten the validity and fairness of the scores they assign to a test taker. Raters are usually affected by their prior experiences and personal backgrounds as they select, weigh, and integrate information into a final judgment. Raters' performance and how they come to a decision about a specific score has been subject of different studies. Although the literature is replete with studies which

quantitatively investigate different predetermined criteria influencing how raters rate, very few studies have specifically investigated the raters' cognitions in terms of the factors they attend to and are aware of in rating a group oral task. A strong need is felt for in depth data-driven studies to tap on the true features that raters attend to in group oral assessment. Trying to fill this void in the literature, this study attempts to underpin factors that influence and account for raters' performance in group discussion tasks. That is, in assigning scores what factors they attend to; what features of the speech sample influence or impress them. An ignorance of such factors may lead to a limited and limiting description of group oral task specificities which may present a construct underrepresentation threat. This inadequacy may be reflected in a reductionist rating scales or inefficient rater training programs. As such, the main objective of this study is to identify the factors that may have been under represented in the literature as actually influencing the raters that may result in inflation or deflation of scores in a group discussion task. A qualitative approach to data collection and analysis may serve this purpose.

So many such factors have been mentioned in literature as accounting for the scores the raters assign to oral tasks. Linguistic features have been amongst the very first factors attended to by researchers as factors influencing the raters in assigning scores in different oral tasks. Hence, a complementary objective of this study is to see to what extent actually the linguistic features of the speech samples influence the raters' perception of proficiency of a learner and assigning a score accordingly. By linguistic features of the speech samples we mean fluency, accuracy, complexity, and amount of talks which are among the factors that are commonly mentioned in the literature as oral proficiency measures. This quantitative phase, intends to see to what proportions, a set of predetermined linguistic features can account for the scores assigned. There is a possibility that other features other than linguistic ones may influence the raters. A correlation between the linguistic features and the scores assigned can fulfil this objective. Using a mixed methods design can present a more comprehensive picture of oral group ratings.

In line with general objectives, the following research questions specifically guide this study:

1- What is the relationship between the scores assigned by the experienced raters and the linguistic features (complexity, accuracy, fluency, and amount of talk) of the speech samples produced by language learners in group oral discussion tasks?

2- What factors do the untrained raters attend to in rating speech samples?

Method

Design of the study

The concurrent triangulation mixed methods design is used to serve the purpose of this study (Creswell & Plano Clark, 2007). Both qualitative and quantitative data were simultaneously collected to enable the researcher detect the factors influencing the untrained raters in group oral assessment. This approach has the advantage of providing 'well-validated and substantiated findings' (Tashakkori & Teddlie, 2003, p.229).

Participants

This study had two different groups of participants. The first group of the participants of this study were 16 Iranian English language learners. They were TEFL students ranging from 19 to 24. Language learners of both genders were selected based on their voluntary participation.

The second group of the participants were 32 untrained Iranian raters who were largely English language teachers of language institutes. Generally, in Iran, language teachers do not receive any formal training on rating. Hence, if the need for rating arises, the language teachers resort to their own experience, background knowledge or rational judgement. The participants were of both genders and varying teaching experience in rating and teaching. In line with the varying years of teaching experience, the raters also varied in terms of age, ranging from 21 to 47. Attempt was made to include teachers of similar education level, namely bachelors, in order to avoid the contaminating effect of education level. Since the raters had to take time listen and rate the speech samples, they were selected based on their voluntary participations and were also paid for the ratings they did.

Instruments

Rating sheet

Listening to the audio files of the group discussion speech samples, the raters assigned each learner a score. They were also required to provide some comments on the rating sheet, pointing to the factors that they attended to while assigning the scores.

Interview

The researchers also had a semi-structured interview with each rater, separately, right after the rating session had ended.

Group oral discussion task

The task implemented in this study was group discussion. The language learners were randomly assigned to the groups of four. Separately, in each group, the participants were supposed to have a discussion over the topic 'early marriage vs. late marriage'. This topic

was considered to be general, familiar, and at the same time interesting enough to the Iranian students and their culture to be discussion rising. No intrusion in the discussion process was made by the researchers and the participants themselves directed the discussions. The discussion took about 15 minutes. The speech samples produced were audio-recorded for further analysis.

Data collection and procedure:

Having collected the speech samples of the learners in the form of group discussions, the researchers asked the untrained Iranian EFL raters to rate them. No training, rating scale or analytical framework was presented to the raters. Listening to the audio files, the raters were supposed to assign each language learner a score of one to six; reflecting basic, elementary, intermediate, upper intermediate, advanced, and mastery levels delineated in the Common European Framework for Reference (CEFR). They were also required to write some comments delineating why they assigned such scores, and what factors they attended to. To avoid order effect of rating, the group discussion speech samples were randomly presented to the raters. After the rating sessions ended, the researchers interviewed the raters individually and their responses were audio recorded. Through repeated careful listening, the researchers transcribed the comments and the interviews. All utterances were written down including both verbal and non-verbal ones like pauses, laughter, pause fillers, etc. Overlaps, repetitions, and false starts were also included.

Data Analysis:

Qualitative:

Content Analysis (CA) was used as the main analysis approach to extract and codify both relevant common and idiosyncratic ideas in the comments and interviews, reflecting the features that the raters attend to in rating which may account for the scores they assigned.

Quantitative:

The linguistic feature measures of the group discussions were also estimated and for each learner an index of fluency, accuracy, complexity and amount of talk were identified to be correlated with the scores the learners received. Spearman rank order correlation coefficient was administered to estimate the correlation between the scores assigned by the raters and the linguistic features.

Analysis of linguistic features: Analysis of linguistic fluency was measured by the rate of speech and quantity of unfilled pauses, which have been found to be significant markers of

fluency (Lennon, 1990; Riggenbach, 1991). For the “speech rate” index, all understandable English syllables, including repeated words and false starts were counted, while non-lexical fillers, such as “um” and “er”, were excluded. This figure was divided by the turn’s time and multiplied by 60 to arrive at the rate per minute (Towell et al., 1996). Unfilled pauses of one second or more within a long turn were timed, and this figure was divided by turn time to give a “pause proportion” index, which was a measure of breakdown in fluency (Tavakoli & Foster, 2008). Amount of talk was also taken as another linguistic feature of the speech sample. It was defined as the total number of words which could be "a reasonable approximation of the amount of floor time occupied by the candidate" (Davis, 2009, p.377). Syntactic complexity was also measured by the ratio of clauses to AS-units and the average length of utterance, which was calculated as the number of words per AS-unit (Foster & Tavakoli, 2009). An AS-unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordination clause(s) associated with either. (Foster, et al. 2000). Finally, the measurement of accuracy was given by the ratio of error free clauses to total clauses. Errors of syntax, morphology and lexical choice will be counted (Nitta & Nakatsuara, 2014).

To check inter- coder reliability, one of the group discussions was randomly selected and coded by a second rater. Inter-coder reliability was high for all four linguistic features (accuracy: 81, complexity 86, fluency: 78 and amount of talk: 100 for < .01).

Findings

Quantitative phase: The relationship between scores and linguistic features

To answer this question a correlation was carried out between the mean of scores assigned by three of the most experienced raters and the linguistic features of the group discussions produced by language learners. Due to the low sample size, Spearman Rank order correlation was utilized.

As evident in table 1, rate of speech, as an index of fluency, showed a correlation estimate of 0.878 with the scores. The amount of talk had a correlation estimate of 0.892 with the scores. Both were statistically significant with a CI of 99% ($p<0.001$).

Table 1. Spearman Rank Order correlation between the mean of scores and the linguistic features of group discussion speech samples

Spearman rho	correlation coefficient	sig.
Accuracy	-179	.506
Complexity: (the ratio of clauses to AS-units)	.096	.724
Complexity: (the number of words per AS-unit)	.246	.358
Fluency: rate of speech	.878***	.000
Fluency: unfilled pauses	-.301	257
Amount of talk	.892**	.000

Hence, the raters had not attended to more delicate and complex features of the group discussion task like accuracy, complexity**,** and quantity of unfilled pauses. This can be justified by Kahneman's (1973) concept of divided attention, which suggested that many factors determine how much attentional capacity be allocated to each task.

Since there were more than one individual in group discussions, the rater were less concerned with delicate linguistic features like complexity and accuracy, dividing their attention to four learners' oral performance at the same time. Hence, they attended to a fewer number of criteria. They also attended to more easy to perceive factors like rate of speech and amount of talk. This may suggest that the raters were not just concerned and influenced by the linguistic features. There may have been other factors that they attended to in a group oral task and which may account for the scores they assigned. The qualitative phase, below, sheds some light on such factors.

Qualitative phase: Factors considered in rating by the Iranian untrained raters

Analysing the data, several emerging patterns reflecting the factors that the raters attended to in assigning scores emerged as described below:

Linguistic features: A qualitative analysis of the data –as well- revealed an awareness on the part of the raters about the linguistic features of the group discussion speech samples while assigning scores. Some of the linguistic features were easier for them to attend to and consider in rating and some other less accessible to them.

Most of the comments concerning accuracy were related to pronunciation errors. Intonation, stress and pronunciation of individual sounds were factors that nearly all raters referred to.

Repeatedly pronounce /d/ for /ð/ or /s/ for /ɵ/

Pronunciation errors like 'advantageous' instead of 'advantages' that make problem for meaning

Farsi intonation

Grammatical and lexical errors did receive some attention. However, compared to pronunciation errors, they received relatively a smaller number of comments.

Good choice of words but grammar problems

Persian expressions and idioms translated into English like 'man of living' or 'see the empty side of the glass'

Wrong words use: like 'unsatisfied' or grammatical structures like: 'getting marriage' or 'the best important'

For fluency, the rate of speech was more eye-catching and easier, as a result, receiving more attention, as was corroborated by the quantitative phase. The quantity of unfilled pauses was also pointed in many cases by the raters; however, mostly in extreme case. That is, where a language learner made a lot of pauses that made his or her flow of speech unnatural.

Some other features were not readily accessible to them, hence they might not have attended to enough. Complexity was one such case. Not all raters were caring about complexity as long as the sentences were accurate and fluently uttered. In rare cases where they did attend to complexity, it was reflected in a comment like:

She used beautiful sentences not just simple sentences

Interactional features: Besides the linguistic features referred to above, the raters were influenced by the interaction features. The most repeatedly mentioned ones are presented below.

The degree of participation: A repeatedly mentioned factor which raters referred to as influencing them in assigning high or low scores was participation; the extent that the participants in the group discussion participated in the discussions. This can be taken as a qualitative counterpart of amount of talk which were shown to be significantly correlated with the scores assigned by the raters in the quantitative phase mentioned above. Participation can have different representations: the ability to initiate a turn, take a turn or hold the floor, etc. Much participation will lead to producing a longer and larger number of turns which will help a participant present**s** himself as proficient, hence, receiving a higher score by the raters. The cooperation in the discussion was usually referred **to** as turn–taking by the raters' familiarity with the technical term and cooperation or participation by those who might not be familiar with the technical term. The following excerpts were taken from the comments provided by the raters on assigning each score and some were extracted from

the interviews they had with the researchers depicting how the raters were affected by the participation quantity in the group discussions:

Self-confident enough to participate in the conversation

Takes a short part in discussion doesn't show herself

She spoke very little so I reduce some points

Because she holds the floor for a long time, I assign her a high mark

Spoke more than others

He didn't speak a lot maybe he felt shy. Maybe because he was the only man in the group. But, I have to reduce some scores.

Sensitivity to the speakers' role in the group discussion: The second recurrent theme in the raters' interviews and the comments on scores accounting for the scores they assigned was the fact that they did attend to the way interlocutors act in relation to each other. To put it technically, they were sensitive to the roles the participants took:

Active vs. passive role: Whether a participant had an active role in the group task which can be represented as listening attentively to other interlocutors, developing and commenting on other's generated turns, asking questions, confirmation check, ability to maintain the floor, challenging or convincing others etc. were deemed as positive features by the raters and inflating the scores they assigned. On the contrary, not following the flow of conversation, just mentioning some points, getting interrupted easily, not raising a question or defending one's own opinion etc. were taken as factors that depicted a participant as passive; leading to a reduction in the scores they received.

A point needs to be clarified here. Although not unrelated, taking a passive or active role should not just be taken equal to participating in the conversation or not. The participation is a much quantitatively measurable factor. However, the extent to which the participant is passively or actively engaged in the conversation is qualitatively different from just participating. An interlocutor can produce much language and take the floor just to –sometimes unenthusiastically- mention his/ her own ideas and not attending to what was mentioned or was relevant to the flow of conversation. The key is to be actively and attentively engaged in the flow of the conversation.

Below is an excerpt from one of group discussion samples:

Maryam: Another thing I want to mention it that […] the boys and girls in our ages emm for example we are in 19, 20,21 we are so er sensitive and we decide on […] on our base and sometimes for example we see a boy fall in love with a girl or vice versa […] about

two or three years after that they divorce because they decide on their feeling. Sara: yes I agree with you. Niloofar: I don't agree with early marriage because maybe emm it has some maybe it has some disadvantages maybe […] the individuals want their educational education […] mmm they spend military service and they something like this. ([…]= pause)

In this excerpt, Nilooafar's contribution was not in line with what was previously mentioned by other interlocutors. As if she did not see the task as a group discussion but a series of monologues, each participant forced to say something to avoid silence. She just said something for the sake of receiving a score and not actually responding to her interlocutors. Such examples received comment like:

She was not answering the previous ones' topic or continue what he was saying

Not see or think themselves as group just talk about her own idea

He seems uninterested in the conversation

He just wanted to pass his turn

Others easily interrupt him

Giving short answers

Does not try to convince others

Waiting to be asked questions

Not initiating any turn

No coherent speech just for the sake of saying something not responding to what has been said

Talking with no enthusiasm

On the other hand, taking an active role was also pinpointed by the raters, as represented in comments like:

Tries to discuss in spite of his bad English

Commenting on interlocutors' speech

Listening carefully, asking questions, giving feedback, talking to everybody

Asking others to give him feedback

Corroborating the influence of an active role on the raters' assessment of the group task, May (2011) identified features such as understanding interlocutor's message, responding to partner, working cooperatively, and contributing to an authentic interaction as factors that raters perceived as interlocutors' mutual achievement.

Supportive Vs. dominant role: The raters were also sensitive to the managing role meaning some participants took, trying to lead the discussion and help the conversation

going. In case of any troubles, they tried to handle them. The following comments reflected this awareness:

Brings a topic for others to follow

Good for group speaking/chat

Asking questions to keep the conversation going

Encouraging her partners speak

Handles the conversation

This managing role can be labelled as supportive and contrasted with the competitive role with comments like:

She stops her friends abruptly

She was in hurry to take turns

He raised a question and he himself answered it immediately to hold the floor.

The objections she made were very direct

Defends bravely takes it like a battle of ideas

This managing role was also identified by May (2006) as a feature that raters perceived as important in rating pair discussion tasks. Using retrospective verbal reports to analyse the factors that raters of paired discussion tasks attended to, May (2006) concluded that the raters did take into account the ability to manage the discussion and work together cooperatively in assessing effectiveness which was the most interactional of the criteria.

Galaczi's (2014) also identified three recurring patterns underlying interactive communication, namely topic development, listener support, and turn-taking management. These themes were relatively in line with the interactive features that Ducasse and Brown (2009) pointed to in their raters' orientation to the learners' construction. These patterns included interactive listening and interactional management which were particularly salient to raters. These two studies confirm the orientation of the raters to other factors other than merely linguistic ones. The managing or supportive role sometime was taken negatively by some raters labelling it as authoritative or dominant. There is a delicate differentiating line between being dominant or just managing the conversation. Hence, raters might have different perceptions of this and hence assign different scores accordingly.

For the first theme -participation- the raters who did pinpoint this factor were quite uniform in terms of the scores they assigned. That is, more participation received a higher score and less participation was assigned a lower scores. However, for the other themes – role of the participants- different interpretations were attributed to the roles. That is, one rater

might consider a participant as dominant and reducing some score and another rater might consider the same participant as just managing, thus, assigning a high score.

In the excerpt below, as well, as if Parisa was Azar's teacher, trying to correct her and asking question to make her talk. This was taken by some raters as the supportive role Parisa took in relation to Azar and assigning a high score to Parisa. Still, some other raters considered this as a negative point since they expected a balanced relation in terms of a group discussion.

Azar: yes I experienced it and because in early marriage we are more […] energyful than mmm than
Parisa: the other who married late
Azar: the other who married late. And we are very.. I am more active and and we can express our feeling to each other
Parisa: better
Azar: better
Parisa: and do you don't you have any problem like house, car or […] supplement for the life?
Azar: its depends on the man. I think the man should be educated and should be educated and can support the […] himself and be on her foot
Parisa: be on his foot, become depen independent

However, in the excerpt below, the fact that Farhad was not caring about the proficiency level of the other interlocutors in terms of repeatedly using complex vocabulary that was beyond the level of his interlocutors was considered as not being supportive and hence did not receive a high score.

Farhad: temporary wedlock is something for alleviating of, alleviation of emotions and feelings, I think this the best solution. This is the most orthodox solution to us. What's your opinion? Sima: would you clarify it? I don't understand

And the following comments were made on these piece of data by the raters:

> *He wanted to show off, he just tried to use difficult words and not caring that his partner did not understand.*
>
> *He just tried to use strange words.*

The overall group interaction pattern: Beside the sensitivity to the role taken by each participant individually, some of the raters also made some references to the overall interactional pattern as a group. Whether a group discussion was symmetric or asymmetric, which was much dependent on the quantity of participation and the roles that participants took. For instance, commenting on a group discussion in which one of the participant uttered less than three complete sentence or in case another participant talked too much giving others

no chance to talk, or when two of the participants address each other and not caring about the rest, some raters referred to the asymmetry in terms of quantity of talk, represented in comments like:

This discussion was three sided.

It was like a dialogue than a group discussion

Hence, the raters did attend to the role, interactional pattern, etc. However, taking a specific role or having a special interaction style led to either positive or negative scores.

The first recurrent theme was the extent to which the interlocutors participate in the group discussion in terms of turn taking. All other linguistic features of accuracy, complexity, etc. being equal, the extent to which a participant could initiate a turn or could take a turn was deemed as a strong point and receiving high scores by the raters. This was corroborated with the quantitative result which showed a statistically significant correlation between amount of talk and the scores the raters assigned. The other two main recurrent themes were the roles that interlocutors took and the overall interactional design of the group. Contrary to the first theme which directly influenced the scores the raters assigned, these two factors might or might not lead to a uniform and predictable influence on the scores. Different raters attributed different interpretation or judgement to an interaction pattern or role.

Conclusion

The study attempted to identify the factors that untrained raters attended to in group oral assessment .The findings of the study can be summarized as follows:

First: Quantitative phase: Only two of the linguistic features, namely the rate of speech as an index of fluency and amount of talk, did statistically correlate with the scores. Regarding the amount of talk, this finding is in line with Galaczi (2008). Comparing peer-peer interaction patterns with the scores the learners received, she found that, although not among the very first features to correlate with the score assigned, amount of talk was one of the topic development discourse features correlating more with the scores assigned by the raters compared to lexical and syntactic cohesive links as features of cohesion between turns. In the case of rate of speech as an indicator of proficiency as perceived by the raters, this finding was also corroborated in studies done by De Jong, Steinel, Florijn, Schoonen and Hulstijn (2012) arguing that articulation rate is one of the best measures of speed fluency. Préfontaine, Kormos and Johnson (2016), as well, found that articulation rate along with the

mean length of runs which is similar to amount of talk proved to be the most influential factors in raters' judgments.

Second: Qualitative phase: As revealed by CA, the raters did attend to some –not all- of the linguistic features, but selectively. However, it turned out that in rating group discussions, the raters attended to other -mostly interactional- features specific to group discussion task as well. Such interactional factors included: the degree of participation, the role of the participants as perceived by the raters and the overall interaction patterns of the discussion. This finding supports the studies that have found that in rating paired or group tasks, the raters attend to mostly interactional features like 'working together cooperatively', 'turn-taking management' or 'interactional management' and 'interactive listening' (May, 2006; Galaczi, 2014; Ducasse & Brown, 2009).

Linguistic features are usually deemed as factors that may correlate with the scores assigned by the rater regardless of the number of interlocutors in the task applied. However, ,as evident in the quantitative phase, the case of group discussions are not limited to linguistic features and may need a much broader scope of investigation.

This pieces of research just scratches the surface of rating group discussion as an oral assessment task. However, grounded in the actual data, the findings can help in group oral assessment. Two main implications of this study are rating scale development and rater training. As a new approach to oral assessment, group discussion tasks may beg for their own specific rating scale reflecting the idiosyncratic features of such tasks which may be missing in other ordinary oral assessment tasks. The fact that the raters do attend to a broader set of factors in assigning scores in group discussion tasks justifies avoiding a reductionist approach which only a predetermined set of criteria are set in a rating scale.

The raters participating in this study were untrained. However, the fact that they did demonstrate an awareness and sensitivity to features specific to a group task, opens a window of opportunity to formally train them in how to rate the features related to group oral tasks both systematically and reliably.

A larger number of raters residing in different educational contexts would have provided a wider range of data. Education level, proficiency, and other variables of the raters can also shed some light on the effects of rater variables on the attention to factors specific to group discussion tasks. For the sake of convenience, this study recorded audio file of the learners engaged in group discussions and played back to the raters to rate them. Recording

video files would have enabled the raters to attend to gestures and body language of the learners too.

References

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

Creswell, J., & Plano Clark, V. (2007). Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367-396.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Linguisti skills and speaking fluency in a second language. *Applied Psycholinguistics, 33*, 1–24.

Ducasse, A, M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. In L. Taylor & G. Wigglesworth (Eds), *Pair work in learning and assessment contexts. Special Issue on Paired Interaction*. *Language Testing*, *26* (3) 423-443.

Foster, P., A. Tonkyn, and G. Wigglesworth. (2000). A unit for all reasons: The analysis of spoken interaction. *Applied Linguistics, 21*(3), 354–75.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866–896.

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

Galaczi, E. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 2*, 89–119.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553-574.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27(*3), 379-399.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*, 370–402.

Hilsdon, J. (1995). The group oral exam: Advantages and limitations. In Alderson, J. & North, B., editors, *Language testing in the 1990s: The communicative legacy* (pp. 189–197). Hertfordshire: Prentice Hall International.

Iwashita, N. (2008). Lexical profiles in EAP speaking task performance. *Indonesian Journal of English Language Teaching*, *4* (2), 111-121.

Iwashita, N. (2010). The effect of intensive recast treatments on the long-term development of less salient structures in Japanese as a foreign language. *Nihongo Kyoiku*, *146*,18-33.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

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

Lazaraton, A., & Davis, L. (2008). A Microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly*, *5*(4), 313- 335.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387-417.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, *7*(1), 25-53.

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

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

Nitta, R., & Nakatsuara, F. (2014). A multifaceted approach to investigating pre-task planning effect on paired oral test performance. *Language Testing*, *31*(2), 147-175.

Ockey, G. J. (2001). Is the oral interview superior to the group oral? *Working Papers on Language Acquisition and Education, International University of Japan*, *11*, 22–41.

Préfontaine,Y., Kormos, J., & Johnson, D.E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language*? Language Testing*, *33*(1) 53–73.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes*, *14*(4), 423–441.

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

Sun, H. (2014). Paired and group oral assessment. *Columbia University Academia Commons*. Retrieved from: www.tc.columbia.edu/tesolalwebjournal

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social & behavioural research*. Thousand Oaks, Calif, Sage Publications.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 58* (2), 439–473.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84–119.

Van Lier, L. (1989) Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489–508.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411–440. Doi: 10.1191/0265532206lt336oa

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University.

Appendix

Rating Sheet

Sara.: **Score:**

Comment
……

Maryam.: **Score:**

Comment
……

Niloofar.: **Score:**

Comment
……

Hasan.: **Score:**

Comment
……

Interview Protocol

1. Describe your experience in rating learners in group discussions.
2. Did you have any specific problem in rating?
3. What factors did you attend to in rating?
4. Did you assign a holistic score or have some pre-specified criteria to stick to?
5. Were you consistent in rating or did your criteria change?
6. Is group discussion a valid task for assessing speaking in your view?
7. Were you certain or hesitant in ratings?
8. Did you feel need any scale or training or collaboration with other raters?