INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 8 No. 3, January 2019, pp. 646-656

Available online at:
http://ejournal.upi.edu/index.php/IJAL/article/view/15266

doi: 10.17509/ijal.v8i3.15266

Assessing ESL undergraduate performance in a group oral test: Rater orientations

Tam Shu Sim[1], Zuraidah Mohd Don[2]*, and Chue Siu Kuan[3]

[1] *Department of English language, Faculty of Languages and Linguistics, University of Malaya, 50603, Lembah Pantai, Kuala Lumpur, Selangor, Malaysia*
[2] *Language Academy, Faculty of Social Sciences and Humanities, University of Technology, 81310 Skudai, Johor, Malaysia*
[2] *Department of English language and Literature, Faculty of Languages and Communication, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia*
[3] *Department of English language, Faculty of Languages and Linguistics, University of Malaya, 50603, Lembah Pantai, Kuala Lumpur, Selangor, Malaysia*

ABSTRACT

Group oral testing has attracted increasing interests in performance-based assessment, mainly due to its ability to measure interactional competence. Increasing attention has consequently been paid to the performance of raters, whose orientations have an impact on the scores in view of their role as mediators between performances and scores. Studies have shown that despite training, raters' personal constructs can lead to different assessments. This study examines rater orientations before and after viewing student performance in an oral test, to discover whether raters subscribe to the view that interaction is individual-focused and is mainly a representation of cognitive or within-language user construct, or whether they believe in a social perspective of interaction. Fourteen participants were interviewed to ascertain their personal constructs for assessing group oral interaction, as well as their justifications for rating a videotaped group oral performance using these personal constructs. The findings show that while raters valued a number of qualities including linguistic abilities, their focus was on interaction. This suggests an inclination towards the social interactional perspective, as they seem well aware that successful interaction is co-constructed, and cannot be achieved through individual language ability alone.

Keywords: Co-constructed interaction; ESL raters' orientations; group oral interaction; higher education; testing group interaction

First Received:	Revised:	Accepted:
4 July 2018	18 November 2018	15 December 2018
Final Proof Received:		**Published:**
28 January 2019		31 January 2019

How to cite (in APA style):
Tam, S. S., Mohd Don, Z., & Chue, S. K. (2019). Assessing ESL undergraduate performance in a group oral test: Rater orientations. *Indonesian Journal of Applied Linguistics, 8*, 646-656. doi: 10.17509/ijal.v8i3.15266

INTRODUCTION

Many students in public universities have insufficient communication skills in English to satisfy the requirements of potential employers and meet the challenges of the global, knowledge-based economy of the present century (Ainol, Isarji, Mohamad, & Tunku, 2012). Potential employers are quite specific about the need for good communication skills in English, which they regard as essential in international business circles, and necessary for information sharing and effective communication and interaction. Consequently, universities are encouraged to review their existing English Language curricula to take a more communicative approach to teach, with the focus on the development of communication skills for reasons of employability and career development (Zuraidah, 2015).

* Corresponding Author
Email: tamss@um.edu.my

The change in the approach has important consequences for student assessment. This article examines one form of assessment, namely a group oral test which assesses the ability of test takers to interact with each other in a specific context in order to complete a task. The inclusion of the interactional perspective in language assessment is relatively new (Chalhoub-Deville, 2003), and has stimulated research into the effect of this form of assessment on test takers and raters (see, e.g. Borger, 2014; Nakatsuhara, 2011). Peer-to-peer tests that take into account the interactional aspect and dynamic process of interactions provide a positive washback effect for teaching and learning in the classroom, and encourage classroom interaction in the form of pair work and group work. According to Van Moere (2006), peer-to-peer tests have increasingly been used to assess second language oral proficiency (Van Moere, 2006), because it provides test takers with opportunities for more genuine interaction and enables them to display a range of language functions and "collaborate and support their interactional partner to co-construct the spoken performance" (Jacoby & Ochs, 1995 cited in Ducasse & Brown, 2009, p. 424).

In examining peer-to-peer tests, the focus has been on the test takers, including their performance and discourse (Brooks, 2009), the effect that interlocutors have on each other during an interaction (Brown, 2003), and the relationship between the scores and test takers' traits (Nakatsuhara, 2011). However, scores also depend on the rater's interpretation of that performance (Papajohn, 2002). Potential rater variability makes it essential to investigate how raters reach their decisions, especially as the assignment of scores has important consequences for test takers (Borger, 2014, p. 17). Ducasse and Brown (2009) make a case for investigating the extent to which scores reflect the qualities the raters themselves value in performance. They argue for the need to identify "features attended to by raters when judging the effectiveness of performance" (p.427).

This article explores one aspect of rating judgements, namely rater orientations in assessing the performance of ESL undergraduates in a specific course-related context. Following Ducasse and Brown (2009), the issues to be addressed are what raters focus on when rating oral performances and whether they are more inclined to the cognitive or the social perspective of interaction. The actual research questions are as follows:

1. What are raters' orientations when making judgements about student performance?
    a. What qualities or features do raters associate in principle with excellent performance in group orals?
    b. What aspects of test taker performance do raters take into account in practice when making their judgements?
2. To what extent do raters
    a. value features of interaction over other aspects of language performance, and
    b. view interaction as a co-constructed achievement?

The examination of these issues is motivated by the need to ensure greater reliability in scoring individual performance in group orals, which will, in turn, have implications for the construct validity of assessment criteria and rating scales.

Rating student performance in peer-to-peer speaking tests

Renewed interests in 'performance assessment' (Bachman, 2000) has drawn greater attention to the role of raters in the assessment process. According to Ducasse and Brown (2009), raters occupy a crucial mediating position between the output (candidates' performance) and outcomes (candidates' scores). Despite training, raters still differ in how they assess, and this leads to different scores (Ang-Aw & Goh, 2011). Rating language performance is a complex matter, requiring a fit between raters' own judgements and the rating that they must apply in the assignment of scores "which involves acts of interpretation …, and thus be subject to disagreement" (McNamara, 1996, p. 117). In support of McNamara (1996), Papajohn (2002) highlights the importance of identifying rater biases and reducing them to an acceptable level.

Previous studies have identified some factors that could contribute to rater variability, and these include raters' linguistic background, gender, first language, previous training or experience, and personality fit between raters and takers (see, e.g. Brown, 1995; Reed & Cohen, 2001). An equally important area, which is little explored and which is the focus of this study, concerns rater orientations defined by Ducasse (2008, p. 7) as "features that raters notice." The features of interaction that raters attend to while scoring and how they apply the rating criteria are relevant to the validity and fairness of scores assigned (Nakatsuhara, 2013).

Borger (2014) foregrounds two pertinent issues concerned with the rating criteria: raters' use of 'implicit' criteria not stated in the descriptors or scoring rubric, and raters' holistic judgements based not on the whole range but on selected features, which "may be more or less salient at different proficiency levels" (p. 37). Raters may intuitively use different rating criteria and so come to different decisions, and the task is complicated even when they are given a specified set of criteria to use (Singto, 2012).

Ducasse (2008, p. 3) highlights the lack "of detailed research into the peer interaction construct," which she attributes to the researchers' insufficient knowledge "about the manner in which raters or candidates construe 'interaction'." Raters' views of what actually constitutes successful interaction may influence their assessment of performance (May, 2011). Peer interaction in pair or group tests poses a great challenge to raters because interaction involves collaboration in "the dynamic process of communication" (Kramsch, 1986, p.368), which is

affected by and mutually dependant on the interlocutors, and which is only observable when interaction is underway. The findings from May (2011, p. 128) showing that raters equated interaction as "mutual achievements" supported the idea that "interactional competence is not what a person *knows*, it is what a person *does* together with others" (Young, 2011, p. 430). Subscribing to this social perspective of performance requires the test takers to consider the bearing they have on each other when they interact to complete an assigned task (Jacoby & Ochs, 1995 cited in Ducasse & Brown, 2009, p. 424), and to assess this requires "the development and validation of improved scales – to measure and report on it" (Ducasse, 2008, p. 2).

While the need for interactional competence in L2 is increasingly recognized and requires peer-to-peer interaction to be included in teaching and testing, developing appropriate assessment criteria is not so straightforward. Thus, it is imperative to amend the "construct of individual ability to accommodate the notion that language use in a communicative event reflects dynamic discourse, which is co-constructed among participants" (Chalhoub-Deville, 2003, p. 373). Young (2011) defines the constructs for tests of interactional competence, but the generalizability of test results remains a challenge.

Recognizing the social dimension in oral proficiency testing involves acknowledging that all test takers are responsible for contributing to and managing the interaction. This makes it more difficult to assess individual performances (McNamara & Roever, 2006). As Chalhoub-Deville (2003, p. 376) points out "If internal attributes of ability are inextricably enmeshed with the specifics of a given situation, then any inferences about ability and performance in other contexts are questionable." While the overall discourse is co-constructed, the scores are assigned to individuals. This creates a challenge in assessing group orals.

The context of the study

As the ability to communicate in English is seen as an advantage for employability, there has been a growing demand for university courses designed to develop speaking skills in English and prepare graduates for the world of work. This creates an urgent need for valid and reliable tests to assess performance.

The context of this study is an intermediate level university English proficiency course which aims to help students to communicate and interact effectively and appropriately in English in social contexts. There are three methods of assessment spread over 14 weeks: pair and group oral tests, and individual presentation. These are high-stakes tests because a pass is required for graduation.

Group oral tests were included to add another dimension of peer-to-peer interaction to the speaking tasks. The aim is to assess whether students are capable of interacting with each other in English to carry out a real-world task. The teacher-raters are provided with the scoring rubric and briefed on how to conduct the test. Although it is a classroom-based assessment, for purposes of objectivity, each class of students is not assessed by its own teacher but by the teacher of another class.

According to Ducasse (2008), there is little research on student performance in group orals because they are not widely used to measure oral proficiency. This lack of detailed empirical information makes it difficult to develop rating criteria that adequately reflect performance. The ultimate goal of this study is to develop evidence-based rating criteria and scales for the assessment of interactional skills in group orals drawing on features of interaction considered important by experienced L2 raters and their judgement about test-taker performance.

METHOD

Two sets of data consisting of rater discourse were collected through semi-structured interviews which enabled the researchers to elicit immediate clarification and elaboration, thus reducing the need to return to the participants later. Initially, the think-aloud protocol was identified as a possible method for eliciting the key features that reflect raters' orientations as they simultaneously rate (see, e.g. Ducasse & Brown, 2009). However, feedback from the pilot study suggested that the participants might not be very forthcoming in expressing their thoughts, partly due to the novelty of the instrument and their diffidence. Other researchers have used a questionnaire with a Likert scale to collect data which is easily quantifiable (Ang-Aw & Goh, 2011), but does not allow the capture of a spontaneous authentic narrative.

The first set of data concerns the raters' general beliefs about what constitutes excellence, while the second consists of holistic ratings and justifications provided by the raters for a videotaped group oral interaction. Beliefs were elicited to get an insight into rater orientations without priming them. They were not presented immediately with the videotaped group oral to avoid having their responses influenced or restricted by the qualities observed in the performance of the candidates. This 'pre-activity' was also intended to foreground the qualities and features which were used as rating criteria for the assessment of test taker performance in the simulated group oral.

The second set of data consisted of rater judgments based on their own criteria which represent "his/her interpretation of the second language proficiency construct" (Fulcher, 2003, p. 35). This can "ensure that the content of the scales is relevant to the context and meaningful to the raters" (Brown, Iwashita & McNamara, 2005, p. 6), which is important for validating the scale development process (Ducasse, 2013, p. 1168). Raters watched the video without being told what features of performance they should focus on.

Participants

This is a small scale exploratory study involving 14 participants with teaching and testing experience. They were individually briefed on the purpose of the study and given the objectives of the course but not the rating criteria. The selection was based on purposive sampling. Of the 52 teacher-raters identified, only 14 volunteered to participate. The participants had from 2 to over 30 years of English language teaching experience, and all had experience of assessing peer-to-peer oral tests, ranging from 2 to over 30 speaking courses. Only 4 indicated that they had undergone some form of formal training in group oral tests. The rest relied on what they had learned from past experience.

Data collection

The oral task was created by the research team and validated by the course coordinator to ensure that it was comparable to test situations used for the actual course assessment. The four test takers were student volunteers preparing to take the test, having completed the required learning hours. They were provided with a task sheet (see Figure 1) requiring them to speak with each other to complete a task.

TEST SITUATION

You and your friends are continuing a discussion for an upcoming presentation. The topic is as follows:

Text messaging has affected students' ability to communicate effectively in formal situations such as writing of official letters.

In this meeting, you plan to discuss the following:

i. Probable/possible reasons for the issue/problem
ii. The best solution or step to overcome the problem.

You have two minutes to prepare and another 20-25 minutes for the discussion.

Figure 1 Task sheet for the oral interaction test

The discussion largely took place between Students 2 and 4. Students 1 and 3 occasionally made independent contributions or supported Students 2 and 4. While the students discussed the reasons and solutions for the problem, they failed to identify the best solution as required by the test task. The whole interaction lasted 22 minutes.

Following Galaczi's (2008) dyadic interactional patterns, the overall interaction can be considered *asymmetric* since two test takers (Students 2 and 4) are much more active. However, it can also be considered *blended* depending on the perspective from which it is viewed. The interactions observed between Students 2 and 4 can be considered *collaborative* and *parallel* as they do, on occasion build on what each other is saying and at the same time, compete for leadership. Interaction involving either Student 2 or 4 with Students 1 and 3 tend to be *asymmetric*. There was no discernible interaction between Students 1 and 3, as neither conversed directly with the other, and discussions occurred with Students 2 and 4 as intermediaries.

Data were collected in three stages. Participants were first interviewed to identify the qualities they associated with excellent student performance in group oral interaction. The question was "What do you think are the qualities of an excellent student in group oral interaction?" Excellence was used as the benchmark to provide a context within which raters could express their expectations. They gave their views without watching the videotaped performance. The interview lasted from 10 to 30 minutes, and it was conducted strictly without any prompt, but merely aimed to probe for elaboration and clarification.

In the second stage, the participants watched a recording of the group oral in its entirety without pausing and then assessed the students using the criteria they had identified. They were not instructed on how to assess the students or given any criteria. During the interview, they were asked to comment on the features that they attended to when making their decision, and this was undertaken in the manner of a retrospective report with minimal questions and clarifications from the interviewer. No time limit was set. In the final stage, the participants were requested to justify their rating. The whole process was audiotaped, and the discourse was later transcribed verbatim.

Data coding

The transcripts were segmented into idea units consisting of one or more utterances with a single aspect of the event expressed as one idea (Green, 1998). In the first data set, the units were coded according to the

features that the participants felt contributed to excellent performance, and those in the second set according to features that the raters attended to as they rated student performance.

Features identified in the coding process were grouped by theme to form major categories, e.g., *linguistic resources* include pronunciation, vocabulary, grammar, and global linguistic resources. Repeated mentions were not counted, but elaborations and clarifications were. For example, interviewee's discourse in Extract 1 was divided into three idea units:

Extract 1:

She has to show display fluency of speech, in terms of the language use, of course, syntactic structures, grammatical coherence and all, vocabulary range, if she's supposed to be excellent, that means she can use words, ah the repertoire of words should show that she is of a higher range…

Idea units for Extract 1

1. She has to show display fluency of speech
2. In terms of the language use, of course, syntactic structures, grammatical coherence and all
3. Vocabulary range, if she's supposed to be excellent, that means she can use words, ah the repertoire of words should show that she is of a higher range…

Zhang and Elder's (2011) coding framework consisting of the categories *Fluency*, *Content*, *Linguistic Resources*, *Interaction*, *Demeanour*, *Compensation Strategies*, and *Other General Comments* was initially used, but the *Interaction* category was found to be inadequate. A subcategory was developed, namely *Conversation Management* and then two other subcategories were added: *Interactive Listening* and *Non-verbal Interpersonal Communication* drawn from Ducasse and Brown's (2009) themes. The coding framework thus combined Zhang and Elder's coding framework, Ducasse and Brown's themes, and what emerged from the data during the coding process. It consists of six categories: *Fluency*, *Content*, *Linguistic resources, Interaction, Demeanour and Compensation strategies.* Each is divided into subcategories (see Table 1).

Members of the research team participated in the coding process, and an inter-coder reliability rate of 79.4% was achieved. According to Ducasse and Brown (2009), the accepted level of agreement in coding is around 80% of the total data coded. To identify the raters' orientations, the number of mentions was counted for each category.

RESULTS AND DISCUSSIONS

This section presents the results of the research analysis to address the two research questions set out in the introduction.

Raters' orientations

The first question examines the raters' orientations to aspects of performance:

1. What are raters' orientations when making judgements about student performance?
   a. What qualities or features do raters associate in principle with excellent performance in group orals?
   b. What aspects of test taker performance do raters take into account in practice when making their judgements?

1(a) is concerned with general beliefs about what constitutes excellent performance in group interaction, which we refer to as "expected qualities" and 1(b) what the raters took into account in practice when assessing an actual group oral interaction, which we refer to as "rating criteria."

The participants mentioned 198 different qualities associated in principle with excellent performance but actually mentioned 434 qualities when rating student performance. The qualities were sorted into six main categories, *Interaction*, *Linguistic Resources*, and *Content* being mentioned more than *Demeanour*, *Fluency*, and *Compensation Strategies*. Table 1 shows the complete list of categories and subcategories and the number of mentions in the respective subcategories.

The most mentioned qualities are those related to *Interaction*. *Interactive Listening*, *Participation in group*, *Dominating*, and *Creating opportunities for others to participate* were mentioned more when participants related their general beliefs about excellent performance. They expected excellent students to listen to others, to participate but not to dominate, and to know when to invite participation. Four sub-categories (*initiating/summarising/ concluding; filling in the gaps/rephrasing*; *gaze*; and *body language*) were mentioned infrequently as expected qualities, but more frequently during the assessment.

The next most mentioned category is *Linguistic Resources*. Excellent performance includes pronouncing words intelligibly and clearly, using a range of appropriate words with grammatical accuracy and being highly proficient. An interesting observation is that the raters used a common set of rating criteria similar to the qualities they identified in Stage 1.

Content includes (a) expected knowledge in the subcategory *ideas*, (b) logical opinions in the subcategory *relevancy to topic*, and (c) the ability to fulfil the set task in the subcategory *appropriateness of response to context*. During the group oral assessment, *ideas* are manifested in terms of quantity, quality, elaboration, and context, whereas descriptions under the other two subcategories did not show any noticeable differences. Before the viewing, *ideas* were mentioned 12 times but increased to 72 mentions when assessing student performance, which indicates that a premium is placed on content (ideas) in oral interaction.

Table 1. Coding framework and mentions in each category/subcategory

Category/Subcategory	Expected qualities (Mentions)	Applied criteria (Mentions)
1. Fluency		
Total mentions, (%)	10 (5%)	16 (4%)
1.1 Fluency (global)	6	9
1.2 Hesitation and pausing	3	6
1.3 Rate of speech	1	1
2. Content		
Total mentions, (%)	31 (16%)	103 (24%)
2.1 Ideas	12	72
2.2 Relevancy to topic	11	18
2.3 Length of speech	-	-
2.4 Appropriateness of response to context	6	12
2.5 Content (global)	2	1
3. Linguistic resources		
Total mentions, (%)	49 (25%)	101 (23%)
3.1 Pronunciation	14	12
3.2 Vocabulary	11	17
3.3 Grammar	8	36
3.4 Linguistic Resources (global)	16	36
4. Interaction		
Total mentions, (%)	88 (44%)	186 (43%)
4.1 Interaction (global)	5	9
4.2 Interactive Listening	9	10
4.3 Participation in Group	9	40
4.4 Intelligibility of others	1	-
4.5 Conversation Management		
4.5.1 Managing topic coherence	5	6
4.5.2 Turn taking	3	8
4.5.2.1 Dominating	18	11
4.5.2.2 Interrupting	7	5
4.5.2.3 Creating opportunities for others to participate	18	19
4.5.3 Leading/sustaining discussion	1	11
4.5.3.1 Initiating/summarizing/concluding	3	24
4.5.3.2 Filling in the gaps/rephrasing	1	11
4.5.3.3 Accommodating others	4	3
4.6 Non-verbal interpersonal communication		
4.6.1 Gaze	1	10
4.6.2 Body Language	3	19
5. Demeanour		
Total mentions, (%)	19 (9.5%)	24 (5%)
5.1 Confidence	7	11
5.2 Others	12	13
6. Compensation Strategies		
Total mentions, (%)	1 (0.5%)	4 (1%)
6.1 Compensation Strategy (global)	-	-
6.2 Specific Compensation Strategies	1	4
TOTAL (All categories)	198	434

Under the category *Demeanour*, the sub-category of *Others* received more mentions than the subcategory of *Confidence* as the former was a catch-all for attributes other than confidence, such as attentiveness and sensitivity. Participants also noted how students' personalities might affect their performance. They were able to consolidate their perceptions during the assessment as the students displayed many of these attributes which either helped or hindered their interaction.

Table 2 indicates the number of participants who mentioned each category as expected qualities and rating criteria. With the exception of *Compensation Strategies*, more than half of the participants mentioned all the categories. When interviewed about their general beliefs about excellent performance, all participants mentioned *Interaction*, followed closely by *Linguistic Resources* and *Content* (12 participants each). More of the categories were mentioned by the participants when they rated the actual performance.

It is significant to note that the relatively large number of comments about the category *Interaction* (see Table 1) reflects a typical pattern found among all the raters; it is not a case of a few raters concentrating on this category. Surprisingly, although *Interaction* features strongly in both the expected qualities and rating criteria, it is not taken into consideration in the scoring rubric for this test, which is based on three criteria: Task fulfilment (28.5%), Language control,

vocabulary and structures (43%), and Communicative ability, fluency and pronunciation errors (28.5%).

This brief overview of the frequency of mentions of qualities reveals the mental constructs that raters bring with them to the rating process. These qualities were mentioned by the participants of this study without any stimulus, and most of them re-emerged with more mentions and fuller narratives during the assessment of the group oral interaction. Hence it seems that complementing rater expectations with actual assessment data enables a more comprehensive picture of rater orientations to emerge.

Table 2. Mentions of the category by participants

Category	Expected qualities Participants (%)	Applied criteria Participants (%)
Fluency	8 (57%)	10 (71%)
Content	12 (86%)	14 (100%)
Linguistic Resources	12 (86%)	14 (100%)
Interaction	14 (100%)	14 (100%)
Demeanour	9 (64%)	9 (64%)
Compensation Strategies	1 (7%)	3 (21%)

Raters' view of interaction

The results from RQ1 were analysed further to address the second research question: To what extent do raters (a) value features of interaction over other aspects of language performance and (b) view interaction as a co-constructed achievement? The answer to this question reflects the raters' view of interaction, whether it is concerned with individual performance or co-constructed achievement, and this can result in contradictory perceptions of the same performance. The former is mainly concerned with interaction as a representation of a cognitive construct which resides in the individual, and the latter the social aspect of performance which takes into account the bearing participants have on each other during the interaction.

Zhang and Elder (2011, p.40) developed a method of calculating the relative proportion of individual rater's mentions of each category of oral proficiency, and suggested that the most frequently invoked category "is arguably the most salient category for each rater group." Raters for whom individual abilities such as linguistic competency are salient are oriented towards a more cognitive or psychological model of L2 interaction, while those for whom interactional abilities are salient are oriented towards a social perspective of interaction.

Based on the frequency of mentions (Figure 2 and Table 1), the raters view *Interaction* as the most important component of group oral interaction, making up 44% and 43% of total mentions for expected qualities and rating criteria respectively. In comparison, *Linguistic Resources* received only 25% and 23% of total mentions as expected qualities and rating criteria. This suggests that the raters valued the ability to interact more than linguistic ability.

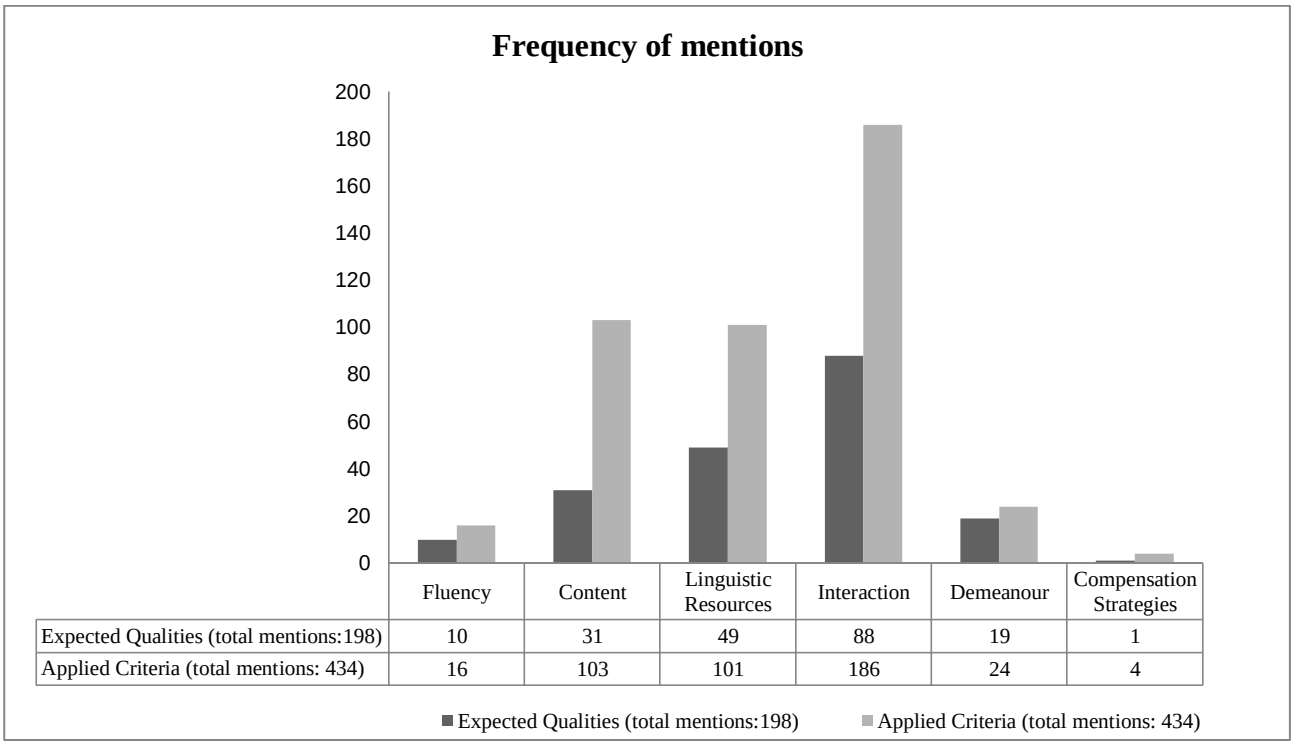

Figure 2. Frequency of mentions according to categories

This study also examined the raters' orientations with respect to other categories. *Fluency* and *Compensation Strategies* (consisting mainly of self-correction and code-switching) are related to language proficiency and were included under *Linguistic Resources*. *Content* and *Demeanour* are associated with *Interaction*. In the case of *Content*, the test situation requires the students to fulfil a task generating ideas or content in the course of the discussion. The raters felt content alone is not enough. The test takers must not only have relevant and good ideas but also know how to contribute these ideas to the discussion. All of these point towards an interactive task fulfilment context. Interestingly, certain qualities such as confidence, sensitiveness, and maturity, which are included under *Demeanour*, were not valued intrinsically but for their contribution to successful interaction. For example, according to one participant, confidence "*is linked with being able to initiate discussion in presenting your own ideas, in a convincing manner and anticipating counter-arguments. So a student who knows all these angles would be able to sell the ideas more convincingly by being confident. It is that level of confidence that allows him to draw the line between leading the discussion and dominating the discussion."*

From this perspective, *Linguistic Resources* accounted for 31% and 28% of mentions of expected qualities and rating criteria respectively, corresponding to 69% and 72% for *Interaction*. These figures suggest that the raters were disposed towards interaction and its peripheral components rather than linguistic ability. The raters also commented that *Participation in the group, Conversation manag*ement which includes *Topic coherence, Turn-taking* and *Leading or sustaining discussion* could only be achieved jointly in interaction with others. This illustrates their orientation towards the social perspective of interaction showing they accept the view that successful interaction involves meaning-making that is jointly constructed.

From the analysis of the participants' discourse it is possible to identify two broad trends with respect to linguistic and interactional abilities: (i) both linguistic and interactional abilities are important, and (ii) linguistic ability is not crucial for successful oral interaction. This would give an indication of the raters' orientation. It is convenient to start with the raters' expectations and then consider the criteria they used when assessing the group performance and justification for their rating.

Both linguistic and interactional abilities are important

Ten raters said that an excellent test taker must have both language skills and interactional abilities. According to Rater H, language proficiency by itself is not enough: "*If an extroverted student has got the language skills …, has all the right ideas but does not know when to stop and when to give others a chance to speak, and in anticipating counter-arguments does not show respect for different views, then [it] does not make him an excellent student. Because the very idea of an oral interaction, for me, is for them to come together to solve the problem, not an individual selling his ideas alone."* Rater K put it more succinctly: "*It's not just speaking, it's interaction*," and according to Rater E an excellent student must "*…have language proficiency,…as well as the other skills to be truly excellent."* However, Rater E also emphasized linguistic accuracy as an indicator of excellence: "[a quality of an excellent student] *would be definitely a very close to accurate use of the language…. And I am talking about close to 96% accuracy."*

The raters specified the criteria that they used to rate student performance and explained their decisions. According to Rater J, language ability and high self-esteem alone did not guarantee good marks; students must participate actively as a group member: *"They are so good* [linguistically and they think] *that's enough to give them a lot of marks and then they will not really interact."* Rater K felt that the students must have the language ability to present their views and know how to contribute to the interaction: "*it just wouldn't be content; it would be also how they deliver the content. They should be able to vocalise their thoughts...and [the student] must be able to place herself within the group as someone who is contributing to the conversation.*" She was prepared to accept some inaccuracy because she believed 'mistakes' were inevitable in speech: "*you can have grammatical mistakes because this is oral…but the basic structures must be there*."

Rater C recognized the importance of linguistic competence, but was of the opinion that a student with interactional ability could still function adequately in group interaction in spite of the lack of linguistic ability: *"*[interactional] *strategies would be helpful and beneficial for the student. Even though the students have poor language proficiency, at least they know how to contribute their opinions despite the grammatical errors or sentence structure."* Likewise, Rater D noted that Student 2 "*shows good interactive skills*" and although "*at times there is a slight groping for certain words, it doesn't sort of interfere with our understanding."*

One rater, however, was ambivalent, while another (Rater E) seemed to be more concerned with linguistic ability and accuracy than interactional skills. The first rater actually drew up two ranking lists instead of one as requested. She could not decide which quality was more important and explained: "*It was very obvious that student 4 tried to get everyone involved; she tried to summarize things for them, and she was very participative...so in terms of strategies she did a good job. But in terms of language proficiency, she scored lower because her sentences are incomplete...quite a few grammatical errors... ."* Although Student 4 was rated first for interactional ability, Student 2 was first for linguistic proficiency: *"Student 2 is slightly better than the other 3 members."*

Rater E commented on the importance of accuracy in communication. Although the group's interactional skills such as turn-taking and topic cohesion were in

general "*not too bad*," and important for successful group interaction, the lack of accuracy affected communication: "*… they lack accuracy; they lack vocabulary, their structures impede actual communication."*

Linguistic ability is not crucial for successful oral interaction

Four of the fourteen raters held the view that oral interaction could be successful even without high language ability. Two did not mention linguistic ability as a requirement. Rater A regarded both interactive abilities and 'knowledge' (i.e., content) as equally important, whereas Rater L only attended to interactional skills, which also included interactive listening. Commenting on the "expected qualities" of an excellent student, Rater L said even without adequate linguistic ability a student could still be excellent if he had all the qualities that she considered important. She stressed that correct grammar and tenses "*don't actually matter to getting the message across because I think that student, with enough encouragement and appreciation, would work to improve his or her grammar...it doesn't mean that just because your grammar is not that great, that you don't have other qualities, that you don't have the cognitive processes…As long as the message gets across and in a manner that everyone is at ease with, and he himself is confident of, then I think that's an excellent student."* Rater N was concerned only with pronunciation: "*... turn-taking is going to be a bit slow probably because one student has said something wrongly or is not so clear in terms of pronunciation. ...[unclear pronunciation] actually distracts the smooth flow of the interaction."*

After the video assessment stage, Rater A commented that all the candidates had made grammatical errors: "*... of course all of them have grammatical errors…The more you talk, the more you make grammatical errors, isn't it?"* When rating student performance she felt linguistic ability was not as important as the ability to interact and explained her position: "*For me, group discussion should be informative, interesting, and interactive. I put [student] number 3 and number 1 as the lowest because they have less contribution than the others."* Rater N conceded that the better candidates (Students 2 and 4) did make grammatical mistakes but compensated in other ways. What is more significant is that she distinguishes between mild and serious linguistic errors and explains why the latter should be penalized: *"...if it's a very mild pronunciation error, it's acceptable for me...but if it is confusing the others...maybe the other person responds, 'Oh you mean before, I thought it was now.' That gives me a signal, 'this person has to be penalized because that mistake has caused the other person to be confused.'"*

What is important to note here is that all the participants felt that the ability to interact was essential while the majority implied that linguistic accuracy was not paramount for a successful oral interaction but was only significant if it stood in the way of communication. This is perhaps understandable in the context of the local ESL setting, which is founded on the communicative approach to language teaching and learning. The raters, who are also teachers on this and similar courses, may not see the need to place too high a premium on linguistic accuracy as long as the communicative intention is achieved.

Hence, in response to RQ 2, it is clear that interactional ability is seen by the raters as a salient quality in group oral interaction and that group interaction cannot be carried out successfully solely according to the individual's language ability. If the task or the test had been viewed merely as a means to assess the test-takers' linguistic ability, it would have suggested that the raters were inclined towards a cognitive/psycholinguistic representation of L2 use. The raters' emphasis on the ability to interact, and the features of interaction they highlighted (such as turn-taking), shows that they regard interaction as a co-constructed achievement since these cannot be expressed individually.

CONCLUSION AND IMPLICATIONS

While the results suggest that the raters believe the ability to interact is important, there is little to indicate their understanding of interactional competence. Further follow-up research is required to provide deeper insights. Secondly, since the study was based on one videotaped interaction, the raters' observations and comments are likely to be limited to the phenomena observed in that particular performance. The data is therefore not necessarily representative because other interactions could yield different observations and interpretations. The small sample size and the specific context of the study limit the conclusions to be drawn, and it is not possible to make generalisations about other testing contexts. Nevertheless, this study has provided some insights into the features and qualities valued by raters in a group oral, and so provided a good starting point for the development of appropriate rating criteria and scales to assess student performance in this particular course or any other speaking courses at a university.

The findings indicate evidence about aspects of performance and features of interaction that raters focus on when assessing student performance in group oral tests. They are similar as well as different from the findings of other studies. The qualities that the raters identified parallel those cited by the participants in Zhang and Elder's (2011) study, from which the coding framework of this study was adapted. Although they found that judging by unguided holistic scores, there were no apparent differences in the consistency or severity of each rater group (native and non-native speakers of English) when judging oral proficiency, in terms of analysis of raters' comments, it was found that Linguistic Resources was mentioned more often by non-

native speakers than native speakers as a relevant factor in their assessment. However, in this study, the majority of the raters who are all non-native speakers of English did not consider linguistic ability to be more relevant than other categories. Like the native speakers in Zhang and Elder's study, they seem to focus on a wider range of abilities in judging oral performance. This suggests that the raters of this study tend to put more emphasis on the test-takers' ability to carry out successfully a real-world task than on the test-takers' underlying language ability as manifested through a particular task.

Similarly, many of the interactional features, such as interactive listening, conversation management, seeking clarification, or extending/sustaining a discussion, mentioned by the raters in this study are not new or unique. They mirror those mentioned in Brooks (2009), Ducasse and Brown (2009), and May (2011). The studies cited were based on paired interaction, while our study looks at group interaction with four speakers. Like the raters in May's study, the participants in this study also support the view that interaction is co-constructed as they too see interaction as a mutual achievement.

One of the major implications of this study would be how the raters' orientations would impact their rating behaviour. Some of them seem to view interaction as more of an individual capability, as evidenced in some of their comments. This could be due to the current practice of having to grade the students individually within an interaction and not as a whole group, thereby making a stronger case for May's (2009, p. 419) argument that a "shared score" should be assigned for "interactional effectiveness" in the test task to acknowledge the co-constructed nature of the interaction. The orientations of the majority of the participants, however, lie within the social perspective of interaction as evidenced from their perceptions that linguistic ability alone is not sufficient for interaction as well as their awareness that co-construction underlies interaction. Given that scores awarded by raters reflect the qualities they value, how will they react to a rating scale which does not mirror their orientations? As mentioned earlier, most of the scoring rubrics used locally, including for this particular course, give more weighting to linguistic proficiency. This is not surprising given that language testing has a history of defining the L2 construct in cognitive terms. By focusing on internal or cognitive abilities, test designers have been able to generalise across contexts, which has been a desired outcome of testing. In this case, considering the fact that the participants have placed a premium on interaction whereas the scoring rubric does not, this mismatch could lead to a loss in test reliability.

Group oral interaction is gaining popularity for assessing oral skills, especially in the ESL context, as it is a "resource-economical" way of assigning speaking scores to large numbers of candidates (Van Moere, 2006, p. 412). The spoken skills portion of the Malaysian University English Test, a high stakes test which has to be taken by all students aspiring to enter public universities, is assessed by means of group oral interaction. As the demand for spoken English increases, so does the demand for a reliable and valid assessment. Given that this study provides further evidence of how raters, who are in a crucial mediating position between test-taker input and test-taker score, highly value interactional qualities, test designers will have to pay greater attention to raters' orientations in developing appropriate rating scales to ensure construct validity and scoring reliability.

ACKNOWLEDGEMENTS

This work was supported by the University of Malaya Research Grants (Project number RP014A-13SBS and RG331-11 HNE. The authors would also like to thank Dr. Mohana Nambiar and Ms. Kan Ngat Har for the significant contribution to the project.

REFERENCES

Ainol, M. Z., Isarji, S., Mohamad, S. N., & Tunku B. T. A. (2012). *English language competency for entry level employment: A market research on industrial needs.* Cyberjaya: Prestariang

Ang-Aw, H. T. & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national level oral examination tasks. *RELC, 42*(1), 31-51. doi: 10.1177/0033688210390226

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42. doi: 10.1177/026553220001700101

Borger, L. (2014). *Looking beyond scores: A study of rater orientations and rating of speaking.* Gothenberg: University of Gothenberg

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366. doi: 10.1177/0265532209104666

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15. doi: 10.1177/026553229501200101

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25. doi: 10.1191/0265532203lt242oa

Brown, A., Iwashita, N & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic Purposes speaking tasks.* Educational Testing Service: Princeton

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369-383. doi: 10.1191/0265532203lt264oa

Ducasse, A. M. (2008). *Interaction in paired oral proficiency assessment in Spanish*. (Unpublished PhD thesis). University of Melbourne, Australia.

Ducasse, A. M. & Brown, A. (2009). Assessing paired orals: Raters' orientations to interaction. *Language Testing, 26*(3), 423-443. doi: 10.1177/0265532209104669

Ducasse, A. M. (2013). Such a nice gesture: Paired Spanish interaction in oral test discourse. *Journal of Language Teaching and Research*, *4*(6), 1167-1175. doi: 10.4304/jltr.4.6.1167-1175

Fulcher, G. (2003). *Testing second language speaking.* London New York: Longman.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the first certificate in English examination. *Language Assessment Quarterly, 5*(2), 89-119. doi: 10.1080/15434300801934702

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge, UK: Cambridge University Press.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366-372. doi: 10.1111/j.1540-4781.1986.tb05291.x

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421. doi: 10.1177/0265532209104668

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145. doi: 10.1080/15434303.2011.565845

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension.* Malden USA: Blackwell Publishing.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508. doi: 10.1177/0265532211398110

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt: Peter Lang

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233. doi: 10.2307/3588333

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin K (Eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Singto, S. (2012). Raters' orientation in a paired speaking test. *Journal of English Studies*, *7*, 139-182.

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

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426-443). New York: Routledge.

Zhang, Y. & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. doi: 10.1177/0265532209360671

Zuraidah, M. D. (2015). *English language education reform in Malaysia: The roadmap 2015-2025.* Putrajaya: Ministry of Education.