Paired and Group Oral Assessment

Haimei Sun[1]

INTRODUCTION

The assessment of second language (L2) speaking has long been an important yet challenging area of research in language testing. L2 testers are often concerned with designing authentic speaking tasks that resemble real-life speaking activities so that score interpretations are generalizable to non-test contexts. The conversational nature of speaking skills has promoted the widespread integration of more authentic and interactive assessment tasks, such as paired or group orals. Such direct test formats typically "involve candidates interacting together to perform a task while one or more examiners observe their performances and rate their language proficiency" (Van Moere, 2013, p. 1).

The earliest attempt to incorporate paired interaction traces back to the Foreign Service Institute Tests (Fulcher, 2003), and to Folland and Robertson (1976) who were first to suggest using group discussion in oral assessment (Fulcher, 1996). In recent decades, the paired or group speaking format has been incorporated within a few large-scale high-stakes tests. For instance, group oral assessment has been integrated within the English A/S level Examination in Hong Kong since 1994 (Swain, 2001) and the College English Test-Spoken English Test (CET-SET) in Mainland China since 1999 (He & Dai, 2006). The most influential adoption of the paired speaking format comes from the University of Cambridge Local Examinations Syndicate (UCLES), who first introduced the paired speaking format in the First Certificate of English (FCE) examination in 1996 (Saville & Hargreaves, 1999; Taylor, 2000).

Given the increasing popularity of paired and group orals, Foot (1999) lamented that little empirical evidence was available to support the use of such a test format. More specifically, he cast doubt on the quality of test takers' performances when communication breaks down due to factors, such as anxiety, different accents, proficiency levels, and personality. Additionally, he criticized that the presence of an interlocutor and an assessor "threatens…the illusion of a natural conversation" (p. 39). His criticism is also concerned with the length of the test. Specifically, he contended that given the same length of test time, paired candidates appear to have far less time to allow adequate amount of linguistic output than in one-to-one interviews, thus challenging any inferences drawn from such limited linguistic samples.

Although Foot's (1999) concerns are not completely unwarranted, due at least in part to the scarcity of empirical research available back then, more researchers argued for the potential benefits of using paired and group interactions to assess L2 speaking. For example, it has been suggested that testing candidates in dyads or groups lowers their communicative anxiety and stress (Ikeda, 1998; Norton, 2005; Saville & Hargreaves, 1999), allowing them to demonstrate their best language proficiency and interactive skills. Furthermore, unlike the traditional oral

[1] Haimei Sun is currently an EdM candidate in Applied Linguistics at Teachers College, Columbia University, where she received an MA in Applied Linguistics. She is also secretary of the TCSOL program at Teachers College. Her scholarly interests center on second language acquisition in general and task-based language teaching and learning, and second language reading in particular. She can be reached at hs2700@tc.columbia.edu.

interview, which is often under attack for eliciting asymmetric spoken discourse, resulting from the unbalanced power relationship between an examiner and an examinee (Johnson, 2001; van Lier, 1989; Young & Milanovic, 1992), paired assessment provides a platform where test takers can produce a wider range of conversation management skills (Galaczi, 2004, 2008; Taylor & Wigglesworth, 2009). Paired assessment also seems to better align with teaching practices, which often attempt to enhance language learning through pair work or group discussion in the classroom, thus engendering positive washback (Ducasse & Brown, 2009; Van Moere, 2013; Saville & Hargreaves, 1999). Other possible advantages of such a test format include cost effectiveness and time efficiency in administration (Ducasse & Brown, 2009) and test fairness with each candidate scored by two examiners (Saville & Hargreaves, 1999).

The debate over the extensive use of paired and group oral assessments has yielded an enriched body of theoretical inquiries and empirical investigations, marking a shift from viewing L2 speaking ability as residing in the individual to emphasizing joint construction distributed among interlocutors within local contexts. This paper aims to review this line of research to provide an enhanced understanding of peer-to-peer interaction so as to better inform and advance L2 speaking construct conceptualization, test design, rating scales development, and rater training. In the remainder of this paper, an account of the theoretical background underlying paired and group orals is provided, and then related empirical studies are reviewed category by category. Thereafter, reviewed research findings are discussed, critiqued, and synthesized. Finally, the paper is concluded with implications and recommendations for future research.

THEORETICAL BACKGROUND OF PAIRED AND GROUP ORALS

The increasing incorporation of paired and group oral assessment in both large-scale and small-scale assessment contexts signifies a focus on the fundamentally social dimension of interaction in second language speaking. This social perspective on interaction differs from interaction in the communicative language ability model (Bachman, 1990; Bachman & Palmer, 1996), which is predominantly cognitive and psycholinguistic in orientation (Chalhoub-Deville & Deville, 2005; McNamara, 1996). Kramsch (1986) first coined the term *interactional competence*, with its theoretical underpinning often drawn from Vygotsky's (1978) sociocultural theory (Chalhoub-Deville, 2003; Fulcher, 2003; Ikeda, 1998), which claims that "[T]here is no universal competence. There are only local competencies, which are situated in a variety of social, cultural, and institutional settings" and that local competence is "acquired through a process of social interaction" (Johnson, 2001, p. 195). This echoes Kramsch's (1986) argument that "successful interaction presupposes…the construction of a shared internal context…that is built through the collaborative effort of the interactional partners" (p. 367).

The notion of interactional competence was later taken up and explicated by He and Young (1998), Young (2000, 2008, 2011), Hall (1993, 1995), and Johnson (2001). According to Hall (1993), interactional competence, which she termed *interactive practices*, is "socioculturally conventionalized configurations of face-to-face interaction by which and within which group members communicate" (p. 144). Extending upon Hall's interpretation, Young (2008) used *discursive practices* to define interactional competence as "a relationship between the participants' employment of linguistic and interactional resources and the contexts in which they are employed" (p. 101). Specifically, those resources include (1) identity resources (i.e., participant framework); (2) linguistic resources (i.e., register and modes of meaning); (3)

interactional resources (i.e., speech acts, turn-taking, repair, and boundaries of the opening and closing acts) (Young, 2011, p. 429-430). Young highlighted that the use of these resources is highly context dependent, therefore varying across different discursive practices.

The above theoretical exploration teases out two fundamental concepts of interactional competence that are of special relevance to paired and group oral assessment. One is *co-construction*, which lies at the core of interactional competence (Fulcher, 2003; Johnson, 2001), and is defined as "the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality" (Jacoby & Ochs, 1995, p. 171). In other words, the co-constructive nature of interactional competence is not an individual attribute, but is shared by all the interlocutors involved in the communication. The other crucial notion is that this co-constructed interaction is *local* (Young, 2011), which indicates that interactional competence is inherently unstable and subject to the influence of the specific social and cultural context in which the interaction takes place. In the context of paired and group speaking assessment, test takers employ their identity, linguistic, and interactional recourses to jointly accomplish a speaking task. The types of resources each individual brings into the interactive communication, including personality, interlocutor familiarity, and proficiency, may well have an impact on their interactional processes and jointly constructed outcomes.

Although the social interactional competence approach appears to offer a more enlightened and nuanced perspective on second language speaking ability, it raises concerns as well when it comes to measurement (Chalhoub-Deville, 2003; Fulcher, 2003; McNamara, 1997; Swain, 2001). Given the essential co-construction of interactional competence, how to disentangle test takers' joint performance to assign each participant an individual score without bias remains an acute issue on L2 testers' research agenda. McNamara and Roever (2006) expressed a similar concern that institutions require individual scores for making informed decisions rather than a faithful account of the interaction. To complicate the matter even further, the emphasis on the local nature of the interactional competence poses a significant threat to the issue of its generalizability. Pressed by these unresolved issues, L2 testing researchers have looked beyond their own field to employ more qualitative methodologies by examining turn-by-turn interaction, so that interactional features are extracted for conceptualizing L2 speaking construct and rating scales. Contextualized in the above theoretical underpinnings, the following section reviews empirical studies of close relevance to paired and group oral assessment, which fall roughly into the categories of test taker characteristics, features of interactional dynamics, and raters' orientation toward co-constructed interaction.

REVIEW OF EMPIRICAL STUDIES ON PAIRED AND GROUP ORALS

Test Taker Characteristics

Of the multitude of test taker variables, the impact of proficiency, personality, and interlocutor familiarity on peer-to-peer interaction appear to have received the most extensive attention. Focusing primarily on L2 learners' levels of proficiency, Iwashita (1998), Csepes (2009), and Davis (2009) investigated the impact of varying interlocutor proficiency levels on test takers' speaking performance. Iwashita's (1998) study involved 20 native speakers of English learning Japanese as a foreign language at an Australian university. The participants

were all females in their 20s and were divided into a high and a low proficiency group. Each participant performed parallel speaking tasks for about 30 minutes, firstly with an interlocutor of similar and then different proficiency level. Two experienced raters scored all the participants' performances on a four-point analytic scale. The data were also transcribed and analyzed in terms of C-units to measure the amount of talk. The results showed that both high proficiency and low proficiency candidates received higher oral scores and produced more output when paired with a high proficiency interlocutor.

Csepes (2009) also investigated the effects of paired proficiency levels on peer-peer performance. Thirty core participants were first selected and then paired with interlocutors of high, low, and similar proficiency, resulting in a total number of 120 participants. The participants were Hungarian students from the same secondary school who were comparable in terms of age and educational background. Although the participants' familiarity variable was controlled, their gender and personalities were not controlled due to practical constraints. The pairs performed parallel speaking tasks for about 5 minutes, which were audiotaped and rated by two trained raters. The findings showed that there was no significant difference between core participants' performance ratings elicited from pairing with different partners of varying proficiency levels. In other words, a core candidate's spoken production was not adversely affected by interacting with a lower proficiency partner or improved by pairing with a high proficiency candidate, thus suggesting no systematic variation in the core candidate's performance. This finding contradicts that of Iwashita's (1998) study.

In view of this controversy and to lend further evidence to the role of interlocutor proficiency, Davis (2009) endeavored to use multifaceted Rasch analysis to examine proficiency effects on test takers' performance, the amount of their production, and its relationship with assigned scores. Twenty-eight mandarin-speakers majoring either in English or software design were recruited to represent two different English proficiency levels. Each participant was paired with a partner from both the same and different major respectively and was instructed to perform a monologue and a paired task. They were given one minute to study the prompts, and their performances were audio and video recorded. Three native speakers of English, who did not receive any formal rater training, scored the examinees' performances by using a 5-point scale rubric. The results revealed that the interlocutor proficiency levels appeared to have little impact on the participants' overall scores, corroborating Csepes' (2009) findings. However, in terms of the quantity of words produced, candidates with a lower proficiency level produced 35% more words when paired with higher proficiency candidates.

Comparing studies on proficiency effect shows both some consistent and inconsistent results. As highlighted above, in contrast with Iwashita's (1998) finding, Csepes (2009) and Davis (2009) found that pairing candidates with partners of differing proficiency levels did not result in significant differences in candidates' speaking scores. This is probably because the mean score differences found in Iwashita's study was not examined for statistical significance. Both Davis (2009) and Iwashita (1998) found that low proficiency candidates produced a larger amount of output when interacting with a high proficiency partner; however, Iwashita emphasized that more output did not necessarily contribute to higher scores. There are also potential variables that may have confounded the findings, as noted by Csepes (2009) who explicitly stated that the personality variable might have influenced the participants' performance. Ikeda (1998) also highlighted the "risk of pairing linguistically compatible learners who may be incompatible personality-wise" (p. 93).

Attempting to assess the specific effects of group members' personality, such as introversion and extroversion, on individuals' performance, Berry (2004) administered a personality questionnaire to 163 Japanese university students (i.e., 78 extroverts and 85 introverts). Two trained raters scored the participants' oral discussions independently. Statistical analyses revealed that both introvert and extrovert participants gained higher scores when assigned in groups with a higher mean level of extroversion, whereas the introverts scored even lower when assigned in groups with an introvert orientation. Using a similar data elicitation technique, Bonk and Van Moere (2004) conducted a large-scale investigation on the effects of shyness on group oral tasks. Informed by an existing questionnaire, they created a shyness survey consisting of ten Likert scale items on a four-point scale. The survey was administered to 1055 Japanese college students after they performed a group discussion task. The results showed that, with their proficiency levels controlled, shyer students demonstrated a slightly significant disadvantage in their group oral performance compared with those who were outgoing. This finding is, to some extent, in accord with findings in Berry's (2004) study in that test candidates who were shyer or more introvert tended to affect those group members who shared similar personalities.

Moving forward from Berry's (2004) and Bonk and Van Moere's (2004) studies, Ockey's (2009) exploration of participants' assertiveness and non-assertiveness on individuals' oral performance in groups was more robust and rigid in terms of the experimental design. The study involved 225 Japanese university students who were divided up into four types of groups of four individuals based on their scores from a personality questionnaire: all assertive, three assertive and one non-assertive, one assertive and three non-assertive, and all non-assertive. Potential construct irrelevant variables including participant familiarity and proficiency levels were built into the experimental design in that participants were grouped with unfamiliar members and high proficiency participants were only placed with those at the same level, with the same true of low proficiency participants. A topic related to campus life was chosen and videotaped instructions were given to familiarize the participants with the discussion task, which took about eight minutes to perform. Two trained raters assigned scores to individuals according to a nine-point scale. Ockey (2009) found that test takers' personality affected their group members' speaking scores. More specifically, assertive candidates were awarded higher scores when assessed with non-assertive partners but were assigned lower scores when assessed with assertive partners. To the contrary, non-assertive candidates' scores were not influenced by their group members' assertiveness.

According to Ockey (2009), assertiveness, a sub-component of extroversion, would have a similar effect as extroversion on a test candidate's score in group orals. Nevertheless, findings from Ockey's study contrast those reported by Berry (2004) and Bonk and Van Moere (2004) who unearthed that both extrovert and introvert candidates gained higher scores when placed in extrovert groups. Ockey (2009) speculated that raters might have perceived assertiveness as a positive trait when assertive candidates led non-assertive ones in a group discussion, but as a negative trait when all assertive candidates within one group competed for holding the floor. The inconsistent finding might also have resulted from the experimental designs. Given that Ockey (2009) exerted rigid control over variables such as personality and interlocutor familiarity, it is unknown whether these extraneous variables, particularly interlocutor familiarity, were taken into consideration in Berry's (2004) and Bonk and Van Moere's (2004) studies.

Studies by O'Sullivan (2002), Ying (2009), and Ockey, Koyama, and Setoguchi (2013) represent attempts to investigate the impact of such interlocutor familiarity on dyadic and group

oral performances. Attempting to test the hypothesis that candidates paired with an acquainted partner would perform significantly better than those interacting with a stranger, O'Sullivan (2002) devised three interactive tasks (i.e., personal information exchange, narrative, and decision making) to elicit data from 32 Japanese participants. All test performances were video recorded and scored by trained raters using an analytic scale and a holistic five-point scale. Statistical analyses not only confirmed the acquaintanceship effect but also indicated this effect was more prominent in affecting test takers' linguistic accuracy, although no significant difference was observed in linguistic complexity.

Building on O'Sullivan's (2002) research, Ying (2009) extended the investigation of interlocutor familiarity to group orals. Her study involved 31 Mandarin speakers with similar English speaking proficiency. There were two group configurations: all-stranger groups and mixed groups with two familiar members and an unfamiliar one. The format of the group task was similar to that of CET-SET[2]. Each group was scored on the spot by two examiners. Besides the interlocutor familiarity factor, the study also included many other facets, such as raters and topics; therefore, a multifaceted Rash analysis was performed. For the purpose of eliciting test takers' perceptions of familiarity effect, a questionnaire was administered upon the completion of the group discussion task. Ying's study yielded similar findings as O'Sullivan's (2002) research, suggesting that test takers experienced more challenges when interacting with a stranger than with a familiar partner. However, interestingly enough, the survey results showed that only 20% test takers expressed a preference for interacting with acquaintances, 30% preferred to interact with strangers, and 50% showed no preference.

A very recent investigation on this topic undertaken by Ockey, Koyama, and Setoguchi (2013) compared class-familiar (n = 146) and class-unfamiliar (n = 159) ratings on a group-speaking task in a Japanese university context. Groups of four consisting of classmates and non-classmates watched an instructional video before performing a discussion task, which lasted about nine minutes with one-minute of preparation time. Twenty-two trained raters using a nine-point scale assessed the group oral performances in aspects of pronunciation, fluency, lexis and grammar, and conversational skills. Following on from Ying (2009), Ockey et al. (2013) also used a survey to provide test takers' perspectives on their preferred group candidates.

The survey results indicated that 55% of the test candidates preferred performing the group discussion task with their classmates whereas 11% preferred strangers. This finding differed from Ying's (2009) report, where the majority test takers showed no preference. Ockey et al. (2013) conjectured that test takers' perception of the test as high-stakes in their study might have resulted in such disparate results. Statistical results suggested no significant difference between the class-familiar and class-unfamiliar groups in either their overall scores or scores in each subscale, which were inconsistent with earlier research findings (O'Sullivan, 2002; Ying, 2009) that confirmed the existence of interlocutor familiarity effects. However, as Ockey et al. (2013) cautioned, given the great many unknowns underlying this line of research, it seems insurmountable to pin down nuanced interlocutor familiarity effects without combining other interwoven elements such as test taker personality and proficiency levels.

The above reviewed empirical studies exploring the relationships between test taker characteristics and individual performances in pairs or groups are fundamentally quantitative and seem to portray an oversimplified dichotomous view of these relationships based on statistical significance levels; therefore, little is still uncovered about the intrinsic nature of the interaction that unfolds between interlocutors with differing characteristics. A microanalytic discourse and

[2] CET-SET includes three sessions: warming up, extended discussion, and follow-up questions.

conversation analysis (CA) approach is more likely to provide greater insights into such turn-by-turn interaction. Therefore, the studies reviewed below are more qualitative in nature, thereby engendering more informative results regarding features of co-constructed discourse.

Features of Interactional Dynamics

To obtain an emic perspective on dyadic interaction discourse, Galaczi (2004, 2008) used a CA approach to investigate peer-peer interaction in paired oral assessment. The purpose of the study was of twofold: to identify *conversation management* patterns and examine their relationships with test scores in the subscale of interactive communication. Data consisted of the third part of the FCE[3] exanimation taken by 30 pairs of test takers with different first languages. Audio-recorded paired performances were transcribed, analyzed, and coded for recurring interactional patterns. Galaczi identified three major categories of interactional patterns, which were *collaborative*, *parallel*, and *asymmetric*. The simultaneous emergence of two of these patterns was termed *blend*. She found these patterns to be distinguishable in terms of mutuality, equality, and conversational dominance. Specifically, the study found that test takers engaging in collaborative interactions demonstrated not only high mutuality in topic expansion and development but also high equality in topic initiation and the quantity of talk. However, parallel interactional patterns exhibited "solo vs. solo" interaction in that although participants engaged in topic initiation and development, they failed to develop other-initiated topics, thus featured high equality but low mutuality. Asymmetric interaction patterns involved candidates assuming either a dominant or a passive role. Relating these conversation management patterns to interactive communication scores, it was found that collaborative groups achieved best performance whereas parallel groups the worse, with asymmetric and blended groups falling in between. Galaczi suggested that these findings have direct implications in developing rating descriptors for the subscale of interactive communication.

Also framed in a CA approach, Gan, Davison, and Hamp-Lyons's (2008) research zoomed in on a specific aspect of conversation management – topic negotiation – in a school-based speaking task in Hong Kong. Data were collected from four secondary school students carrying out an eight-minute discussion task on a gift proposal for a character in the film *Forrest Gump*. The discussion was video recorded, transcribed, and coded following a bottom-up and iterative process. Gan et al.'s (2008) qualitative analysis revealed that topicality negotiation ensued with test-takers' clarification of the task demand, which, according to Sacks (1992), functioned as "transitional first" or "false first" topic talk. It was also noted that test takers demonstrated marked topic shifts by using signal moves and stepwise topic movement by referring back to previously mentioned content. The authors maintained that their findings lend further evidence to the potential benefits of paired and group oral assessments in creating an equal exchange system and generating more varied speech functions.

The following two studies (He & Dai, 2006; Lazaraton & Davis, 2008) also examined test takers' discourse features but with slightly different purposes. He and Dai's (2006) study explored the extent to which interactional language functions elicited via group oral discussion matched that specified in the CET-SET syllabus, thus providing empirical evidence for the validity of the CET-SET. To this end, they built up a corpora of test taker' group discussion section of the CET-SET administered in December 2001. The data were coded for interactional

[3] FCE consists of three parts: an interview between the interviewer and a test taker, a monologue task, and a two-way collaborative task.

language functions, specified in the CET-SET Syllabus, including (dis)agreeing, asking for opinions or information, challenging, supporting, modifying, persuading, developing, and negotiating meaning (Ministry of Education, 1999). A questionnaire was also administered to 196 candidates upon their completion of the test. The results showed that only two interactional functions occurred most frequently, (dis)agreeing and asking for opinions or information, which accounted for nearly 75% of the total number of interactional functions elicited. The authors were also surprised to find that candidates tended to concentrate on organizing their own thoughts and used lengthy turns, assuming that their performances were determined by their quantity of production. The test takers also mistakenly believed that their target audience was the examiners rather than their group members.

Lazaraton and Davis (2008) also worked backward to identify discourse features that could match analytical ratings and the scores assigned, and enabled test takers to position themselves as being proficient in speaking tests. The data of the study consisted of videotaped recordings of Cambridge ESOL's FCE and Preliminary English Test (PET) examinations, and were transcribed following CA conventions. By providing turn-by-turn interactional segments, 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). Therefore, the authors argued for fluidity of proficiency, as it changes depending on the interlocutor and the identity resources s/he brings to the interaction, thus indicating interlocutor effects on candidates' oral performance.

Inspired by He and Dai's (2006) and Lazaraton and Davis's (2008) research findings, Luk (2010) conducted a comprehensive investigation of interactional features in a group oral assessment in a Hong Kong school-based assessment context. In addition to pinpointing macro and micro discourse features in interactions, the author also attempted to uncover if these discourse features revealed test takers' desire to present a best self-impression for evaluation purpose. The participants were 43 female secondary students and their course instructor. The students were randomly assigned into groups of four and given six minutes to carry out a group discussion task on a given text prompt. As well as transcripts of group discussions, data were also collected through a questionnaire and interviews with the instructor and six participants. The results revealed eight key features: (1) recurrent frames, types of talk, and speech acts; (2) ritualized opening and closing; (3) orderly turn-taking practices; (4) heavier-weighting and front-loading content delivery frames; (5) frequency surface converging responses; (6) avoidance of negotiation; (7) self-initiation to avoid dead air; and (8) role-playing critical correspondents (Luk, 2010, p. 34-42). The authors argued that these discourse features demonstrated test takers' desire to obtain high scores by presenting themselves as efficient interlocutors rather than engaging in authentic communication.

As well as the above studies to identify discourse features and the underlying identities interlocutors created, there are also studies (Galaczi, 2014; Gan, 2010) that compared specific conversational features employed by interlocutors of different proficiency levels. Gan (2010) compared interaction features between high- and low-scoring groups in a secondary school context in Hong Kong. Test candidates carried out group discussion tasks similar to those described in Gan et al. (2008). Two recordings of students performing group oral tasks were selected, with one representing the higher-scoring group and the other the lower-scoring group. A moment-by-moment analysis of transcripts revealed that test takers from the high-scoring group demonstrated constructive and contingent engagement with each other's ideas and a wide

range of speech functions such as suggesting, agreeing/disagreeing, explaining, and challenging. With regard to the lower-scoring group, although students failed to show contingent topic development, they managed to offer mutual support for each other through prompting and co-construction, thus prioritizing friendly discourse maintenance over idea expansion.

Galaczi (2014) also investigated co-constructed discourse in paired speaking tests across different proficiency levels of the Common European Framework for Reference (CEFR). Her study included test candidates at CEFR levels B1 to C2, but only pairs at each level awarded between band 3-4[4] ratings on the *interactive communication* subscale were chosen, resulting in 41 paired recordings. The data were transcribed following CA conventions and coded for reoccurring interactional patterns. Three general categories were established, which were *topic development*, *listener support*, and *turn-taking management*. Both qualitative and quantitative analyses showed that as the proficiency level increased, extension of both self- and other-initiated topics increased, with the highest level demonstrating strong interactional competence in expanding other-initiated topics and joint construction of discourse. In terms of listener support, this started to emerge from B2 level featuring primarily backchannels and with C1 and C2 levels featuring both backchannels and confirmations of comprehension. Lastly, it was found that the ability to initiate a turn after a latch occurred more frequently as language proficiency levels increased.

It is evident from the above that microanalytic approaches such as CA provide a fine-grained account of the interactional dynamics embedded in paired and groups orals. Furthermore, Galaczi (2004, 2008), Gan (2010), Galaczi (2014) collectively showed that test takers engaging in dyadic and group interaction produced a broader range of conversation management skills and demonstrated different interactional patterns and/or features, with more proficient pairs showing collaborative and contingent development of self-and other-initiated topics, turn-taking management, and listen support. However, the findings that reported the elicitation of underrepresented linguistic functions (He and Dai, 2006) and orderly turn taking among participants (Luk, 2010) probably relate to Foot's (1998) observation that a paired test may not be suitable for low proficiency learners. Interactional patterns related to participants' response to task demand (Gan et al., 2008; Luk, 2010), and identity work in a test situation (Lazaraton & Davis, 2008; Luk, 2010) were also observed. Given all the interactional features identified, it is vitally important to examine if these features are compatible with what raters attend to in their rating, as dealt with below.

Raters' Orientation toward Co-constructed Interaction

Orr's (2002) study represented one of the earliest attempts to explore raters' perspectives in co-constructed speaking assessment. The study aimed to recapture the rating process of the Cambridge FCE speaking test through oral examiners' retrospective verbal reports. Thirty-two FCE examiners rated two video recordings of two paired interviews using an analytical scoring rubric that included grammar and vocabulary, discourse management, pronunciation and interactive communication. Raters' verbal protocols were recorded, transcribed, and coded. The findings revealed that not only did discrepancy persist in raters' interpretation of the rating criteria and rater severity but also in the underlying rationales for assigning the same score to the same candidate. It was also found that raters tended to heed non-criteria features. The analysis

[4] The band ranges from 1 to 5. Pairs awarded band 3-4 ratings are considered average.

highlighted three aspects that the majority of raters attended to when deciding on a score, regardless of the rating criteria. These three aspects were the global impression of a candidate's performance, comparing candidates in pairs, and a candidate's self-presentation.

A similar study by Ducasse and Brown (2009) also explored the interactional features that raters focused on in scoring. Situated within an Australian university context, the study involved 34 beginners learning Spanish as a foreign language, with 12 teacher raters, with all but one being a native speaker of Spanish. The participants, who were familiar with each other, chose their own partner to engage in a paired discussion task on three familiar topics for ten minutes. The paired discussions were video recorded, with each participant assessed by at least two raters. The raters were not given any guidance as to what they should concentrate on in their rating. Raters' tape-recorded comments were transcribed and coded, resulting in three categories, including *non-verbal interpersonal communication, interactive listening*, and *interactional management*. The authors suggested that the saliency of these features to the raters indicated they are important aspects of interactional competence. They also highlighted that *interactive listening* lies at the heart of co-constructed dialogues and confirmed that paired interaction elicited more varied interactional management skills.

The most comprehensive investigation into rater orientation toward paired interaction to date comes from a series of studies by May (2006, 2009, 2011) who not only scrutinized the interactional features salient to raters but also probed into the thorny issue of how to assign individual scores to jointly constructed performances. To pinpoint features of paired speaking performances salient to raters, May (2006) recruited 12 Chinese students, with intermediate to advanced oral proficiency levels, enrolled in an intensive English for Academic Purposes (EAP) course in Singapore. Each candidate was paired with a partner of a similar and then a different proficiency level to perform two discussion tasks respectively. The discussions were based on reading passages the candidates had been given to read beforehand to align with a theme-based high-stakes test. Each pair was given five-minutes planning time and then their discussions were video recorded. Two trained raters assessed six of the paired discussions with an analytic rating scale (fluency, accuracy, range, effectiveness, and overall), producing 12 retrospective verbal protocols. The verbal protocols were transcribed and coded, and frequency was counted and tallied. The results showed that the two raters demonstrated different interpretations of the rating scale and more than 30% of their comments related to non-criteria features of the performance, as found in Orr's (2002) study. May (2006) also emphasized that raters had to constantly "reconcile aspects of complex paired candidate interactions with rating scales and their own frames of reference as both teachers and raters" (p. 47).

Extending her initial exploratory examination of raters' responses to paired candidate performances, May (2009, 2011) carried out two further investigations using the same paired interaction dataset as described above in 2006, but the 2009 and 2011 studies included four trained raters, from whom data, including initial ratings, stimulated verbal recalls, rating notes, and paired rater discussions, were elicited. While the 2009 study focused primarily on raters' orientation toward asymmetric patterns of paired interaction, the 2011 study focused on the operationalization of the interactional competence construct.

Using Galaczi's (2004, 2008) categorization of interactional patterns, including collaborative, asymmetric, and parallel patterns, May (2009) found that only two out of 12 paired speaking tests were identified as asymmetric interactions, which appeared to pose a challenge to raters in assigning individual scores to co-constructed performances, as these were cases where one candidate might have been disadvantaged by the other. It was also revealed that in the case

of collaborative patterns of interaction, features such as mutual comprehensibility, effective responses, and the authenticity and quality of interaction were perceived as mutual achievements by raters, thus meriting the awarding of a shared score. However, May cautioned that interactions in the target language use domain, such as conversing with a professor, may not always turn out collaborative, thereby suggesting using multiple task formats.

May's (2011) study identified features that raters perceived as interlocutors' mutual achievement, which included understanding interlocutor's message, responding to partner, working cooperatively, and contributing to an authentic interaction. She also recommended including nonlinguistic features, such as body language, in operationalizing interactional competence. However, a concern was raised regarding raters comparing candidates' performance against each other rather than the rating criteria.

This review of the raters' perspectives using verbal reports complements this paper's review of empirical inquiries pertaining to paired and group oral assessment. It is perhaps surprising yet illuminating to find that raters attend to non-criteria features, particularly nonlinguistic features, and have a tendency to compare candidates' performances with one another (May, 2006, 2011; Orr, 2002). Interactional features such as interactive listening (Ducasse & Brown, 2009), mutual comprehensibility, and authenticity and quality of interaction (May, 2009, 2011) are particularly salient to raters, therefore shedding light on conceptualizing the construct of interactional competence.

DISCUSSION

The empirical studies in the above section inform three essential issues surrounding the use of a paired and group speaking format: whom one should be paired or grouped with, what prominent features peer-peer interaction exhibits, and how raters approach the co-constructed discourse. These issues are interrelated in that the examination of interlocutor effects (e.g., proficiency, personality, and familiarity) not only provides insights into pairing and grouping candidates but also into explaining the underlying interactional features identified. In addition, identifying these features from both test takers' production and raters' rating process aspects proves especially informative and meaningful in conceptualizing speaking construct, developing rating scales, and providing guidance in rater training.

Studies pertaining to test taker characteristics appear to be more quantitatively oriented in nature, with most studies focusing primarily on one of the following interlocutor variables: proficiency, personality, and interlocutor familiarity and their effects on paired or group speaking scores. With regard to proficiency levels, Csepes (2009) and Davis (2009) found no significant difference in interlocutors' speaking ratings, confirming Lazaraton and Davis's (2008) observation that "various manifestations of the interlocutor effect do not necessarily translate into increased or decreased ratings" (p. 330). In terms of personality, while Berry (2004) and Bonk and Van Moere (2004) revealed that both extroverts and introverts performed better when grouped with extroverts, Ockey (2009) found assertive candidates getting lower scores when interacting with other assertive group members. One possible explanation for Ockey's differing finding is that when assertive candidates competed with each other for turn taking, they might just concentrate on developing their own ideas, thus leading to parallel patterns of interaction as shown in Galaczi's (2004, 2008) studies which found parallel groups receiving the lowest scores. Regarding interlocutor familiarity, O'Sullivan (2002) and Ying (2009) provided empirical

evidence of an acquaintanceship effect on test takers' oral performance, whereas Ockey et al. (2013) did not, probably because it was potentially problematic to operationalize familiarity in terms of classmate/non-classmate in Ockey et al's study, as it was likely that two non-classmates were familiar with each other.

It is always desirable to exert rigid control over extraneous variables in quantitative studies to avoid any confounding effect. However, given the lack of control over extraneous variables in most studies above and the interwoven nature of test taker characteristics, it is not unexpected to yield conflicting results regarding these interlocutor variables. Ockey et al. (2013) contended "Identifying [interlocutor familiarity] effects would not be possible in a study that did not take into consideration student personality/profiles or differing proficiency levels as an experimental element" (p. 304). Also noted from the above studies is that five (Berry, 2004; Bonk & Van Moere, 2004; Ockey, 2009; Ockey et al., 2013; O'Sullivan, 2002) out of the nine studies involved Japanese participants representing a particular cultural group who "are generally not inclined to state an opinion on issues or discuss topics at length with a stranger" (White, 1989, p. 70). Therefore, findings obtained from this line of research, especially regarding interlocutor familiarity and personality, may not generalize to test takers from different cultural backgrounds. It should also be noted that the relationship between interlocutor effects and test takers' speaking performance is not simply dichotomous or static, as it may vary depending on the identity, linguistic, and interactional resources one brings into the interaction, thus reflecting the local nature of the co-construction (Young, 2011).

Unlike research on test taker characteristics, studies investigating interactional features and raters' orientation towards joint construction are more qualitative by employing discourse or conversation analytic approaches, thereby allowing for analyses of turn-by-turn interaction through iterative coding and identification of reoccurring patterns. Identifying interactional features from both candidates' discourse and raters' rating process provides an overarching view as to what is produced and what is actually rated. Analyses of co-constructed discourse and raters' verbal protocols uncovered a few overlapping interactional features, providing insight into the construct of interactional competence and rating scales development. In Galaczi's (2014) most recent study, she identified three reoccurring patterns underlying the subscale of interactive communication, which are topic development, listener support, and turn-taking management. These patterns more or less correspond to what Ducasse and Brown (2009) reported on their raters' orientation to joint construction. They found that interactive features such as interactive listening and interactional management are particularly salient to raters. Another overarching feature is the test candidates' self-presentation or self-positioning. Lazaraton and Davis (2008) and Luk (2010) uncovered interactions where candidates presented themselves as being proficient and interactive interlocutors, which is perhaps in accordance with raters' global impression test takers' performance as shown in Orr's (2002) study. However, there are also interactional features that existed but which raters did not attend to. For example, Gan et al. (2008) and Luk's (2010) studies showed orderly openings and closings related to given task prompts before candidates engaged in real discussion. Studies from the raters' perspective revealed raters paying attention to non-linguistic features and comparing performances between candidates.

With respect to the rating of joint construction, May (2009) discovered that, among those patterns (i.e., collaborative, parallel, and asymmetric) identified by Galaczi (2004, 2008), asymmetric interaction seemed to present a great challenge to raters whereas collaborative patterns did not. Interlocutors engaging in the collaborative pattern exhibit high mutuality and

equality, indicating that features, such as mutual comprehensibility, effective responses, and the quality of interaction, are mutual achievements that entail the awarding of a shared score. However, in cases where asymmetric patterns of interaction occur, raters are often pressed to disentangle the co-construction to assign individuals a fair score. Awareness of interactional features relevant to different proficiency levels might help facilitate rating in such cases. Gan (2010) and Galaczi (2014) showed that high proficiency test takers demonstrate the ability to engage in contingent development of both self- and other-initiated topics, show listener support, and use a wider range of conversational management skills whereas low proficiency test takers tend to focus on self-initiated topics and demonstrate minimal interactive listening. If these key features are incorporated into rating scales and rater training, raters will probably achieve a more consistent construal of rating scales, which therefore helps ensure reliable scoring and valid interpretation of test scores.

Although the micronalytic approach seems to offer a more comprehensive and nuanced perspective on unraveling interactional features embedded in paired and group orals, it is not without drawbacks. One of the drawbacks that L2 testing researchers are most concerned with is the issue of generalizability. Specifically, it is not uncommon to find features derived from one dataset or the presentation of a few extracts not generalizable to different social and cultural contexts. The very subjective nature of data handling in discourse or conversational analyses may also pose a threat to the attempt to achieve objectivity in assessing interactional competence. Furthermore, the local nature of interactional competence adds to the constraints of the generalizability issue (Chalhoub-Deville, 2003). Another possible and more practical drawback is that the use of audio-only recording (Galaczi, 2004, 2008), probably due to practical constraints, may limit the interpretation of research results. This is highlighted in the amount of research (Ducasse & Brown, 2009; May, 2009, 2011; Orr, 2002) on raters' scoring process that unanimously reported on raters attending to non-verbal behaviors. Considering the saliency of such nonlinguistic features to raters, May (2011) even suggested building these features into the construct of interactional competence.

CONCLUSIONS

The empirical research studies reviewed in this paper offer a comprehensive up-to-date overview of the findings generated from both quantitative and qualitative investigations into paired and group oral assessments. Research findings indicate the intricately complicated issue of interlocutor effects on test takers' oral performance, varied interactional features of peer-to-peer interaction, and raters' orientation toward joint construction. These findings lend strong empirical evidence to the theory of interactional competence that emphasizes co-construction of discourse among interlocutors within locally situated social contexts, thus providing L2 testers with a more in-depth understanding in conceptualizing L2 speaking construct. The quantitative and qualitative methodological approach to this line of inquiry elicited some enriched and complementary findings. The finding that raters' inconsistent perception of rating criteria and attending to non-criteria aspects is particularly insightful, suggesting the importance of incorporating the raters' perspective in developing rating scales (Pollitt & Murray, 1996).

As Chalhoub-Deville (2003) claims, the shift to a sociocultural perspective on L2 speaking ability is only a recent phenomenon, thus indicating the need of ongoing research to probe into the nature of interactional competence, its operationalization, and scoring of co-constructed performances. Specifically, in light of current research focusing mostly on Japanese

and Mandarin speakers, future research is expected to recruit participants from varied cultural backgrounds to capture the complex matter of interlocutor effects. It is also highly recommended to investigate more than one interlocutor variable at a time to engender the dynamic relationships between these variables and test takers' speaking performance. For future endeavors that attempt to disentangle the scoring of joint performance, parallel and blended patterns of interaction identified by Galaczi (2004, 2008) are important aspects that still remain unexplored.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Berry, V. (2004). A study of the interaction between individual personality differences and oral performance test facets. Unpublished doctoral dissertation. King's College, University of London, UK.

Bonk, W. J., & Van Moere, A. (2004). L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores. Paper presented at the annual meeting of the Language Testing Research Colloquium, Temecula, California.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20,* 369–383.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815-832). Mahwah, NJ: Lawrence Erlbaum Associates.

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance.* New York, NY: Peter Lang.

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

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26,* 423–443.

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

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

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal, 30*, 156-167.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 70-76.

Galaczi, E. (2004). Peer–peer interaction in a speaking test: The case of the *First Certificate in English.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York, USA.

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.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring

students. *Language Testing, 27*(4), 585-602.
Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistic, 30*(3), 315-334.
He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam and Philadelphia: John Benjamins.
He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*(3), 370-401.
Hall, J. K. (1993). The role of oral practices in interaction with implications for learning another language. *Applied Linguistics, 14*, 145-166.
Hall, J. K. (1995). (Re)creating our worlds with words: A sociocultural perspective of face-to-face interaction. *Applied Linguistics, 16*, 206-232.
Ikeda, K. (1998). The paired learner interview: A preliminary investigation applying Vygotskyan insights. *Language, Culture and Curriculum, 11*(1), 71-96.
Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-65.
Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*, 171-183.
Johnson, M. (2001). *The art of nonconversation*. New Haven, CT: Yale University Press.
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,* 313–335.
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. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*(1)*,* 29–31.
May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26,* 397-422.
May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145.
McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Addison-Wesley Longman.
McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.
McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.
Ministry of Education (1999). *College English teaching syllabus*. Shanghai, PRC: Shanghai Foreign Language Education Press.
Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT, 59*(4), 287-297.
Ockey, G. J. (2009). The effects of group members' personalities on a test-taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186.
Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*(3), 292-308.
Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System,*

30, 143-154.
O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.
Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 74-91). Cambridge, UK: Cambridge University Press.
Sacks, H. (1992). *Lectures on conversation*. Oxford: Blackwell.
Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*(1), 42-51.
Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302.
Taylor, L. (2000). Issues in speaking assessment research. *Research Notes 1*. Cambridge: Cambridge ESOL.
Taylor, L., & Wigglesworth, G. (2009). Are two head better than one? Paired work in L2 assessment contexts. *Language Testing, 26*(3), 325-339.
van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489-508.
Van Moere, A. (2013). Paired and group oral assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-4). Oxford, England: Wiley-Blackwell.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University.
White, S. (1989). Backchannels across cultures: A study of American and Japanese. *Language in Society, 18*, 59-76.
Ying, B. (2009). The impact of familiarity on group oral proficiency testing. *CELEA, 32*(2), 114-125.
Young, R. (2000). Interactional competence: Challenges for validity. Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Vancouver, Canada. ERIC 444361. Retrieved on November 10th, 2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.9903&rep=rep1&type=pdf
Young, R. (2008). *Language and interaction: An advanced resource book*. London and New York: Routledge.
Young, R. (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). London & New York: Routledge.
Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 403-424.