Developing speaking assessment tasks to reflect the 'social turn' in language testing

LYN MAY

Queensland University of Technology

ABSTRACT

Interactional competence has emerged as a focal point for language testing researchers in recent years. In spoken communication involving two or more interlocutors, the co-construction of discourse is central to successful interaction. The acknowledgement of co-construction has led to concern over the impact of the interlocutor and the separability of performances in speaking tests involving interaction. The purpose of this article is to review recent studies of direct relevance to the construct of interactional competence and its operationalisation by raters in the context of second language speaking tests. The review begins by tracing the emergence of interaction as a criterion in speaking tests from a theoretical perspective, and then focuses on research salient to interactional effectiveness that has been carried out in the context of language testing interviews and group and paired speaking tests.

THE EMERGENCE OF INTERACTION AS A CRITERION IN SPEAKING TESTS

The importance of interactional competence was highlighted by Kramsch (1986), who called for a deeper understanding, particularly

Address for correspondence: Lyn May, School of Cultural and Language Studies in Education, Queensland University of Technology, Victoria Park Road, Kelvin Grove, Queensland, 4059; Email: lynette.may@qut.edu.au

University of Sydney Papers in TESOL, 5, 1-30.

 ISSN: 1834-3198 (Print) & 1834-4712 (Online)

in terms of operationalising this construct in speaking tests. Kramsch was strongly critical of the focus of existing tests on lexis and grammar, rather than incorporating the "dynamic process of communication" (p.386). This focus on the individual candidate was particularly noticeable at a time when communicative language teaching pedagogy emphasised the importance of classroom interaction, which often included pair and group work. The issues raised by Kramsch, including the complexity of the construct of interactional competence, the impact of the interlocutor and the inherently shared responsibility for interactional patterns between interlocutors, inspired a research agenda that continues to the present.

The concept of co-construction is integral to interactional competence, and Jacoby and Ochs (1995: 171) define it as incorporating a "range of interactional processes, including collaboration, cooperation, and coordination". The shared responsibility for the success of a discussion is clearly articulated by Jacoby and Ochs, as they identify the "distributed responsibility among interlocutors for the creation of sequential coherence, identities, meanings and events" (p.177). This concept is also reflected in the work of Hall (1995), who defines talk as being composed of "interactive practices" that reflect the complexity and interconnection of a community of speakers. Advocating a sociohistorical perspective on language learning and assessment, she argues that "language use and language learning are not solely individually motivated and unconstrained activities … one's participation is tied not only to who one is, but to the kind of practice one is engaging in, and the degree of conventionality, authority, that is embedded in the meanings of the resources available to one" (p.221). The notion of discursive practices was further developed by Young (2000: 1), who defined a discursive approach to language-in-interaction as manifesting "a view of social realities as interactionally constructed rather than existing independently of interaction, of meanings negotiated through interaction rather than fixed in advance of interaction....". This perspective on spoken interaction

posed difficulties for those involved in the development of direct speaking tests, where a candidate is required to interact with an interlocutor. If discursive practices were only meaningful in the context of a particular interaction, then the issue of the generalisability of the results becomes paramount. Reflecting on the challenges posed to the field by the "social interactional perspective", Chalhoub-Deville (2003: 373) proposed the incorporation of co-construction into the previously cognitively-oriented view of individual competence, and acknowledging the complexities that are inherent when trying to reconcile "the notion that language ability is local" with "the need for assessments to yield scores that generalize across contextual boundaries".

McNamara (1996, 1997) also problematised the assumption that communicative competence resides within the individual, arguing that current models were flawed because they "focus too much on the individual, rather than the individual in interaction" (1996: 85). This concern was echoed by Young (2000: 5), who differentiated communicative competence from interactional competence in that while the communicative competence framework "helps us to understand what an individual needs to know and to do in order to communicate", interactional competence is characterized by the focus on co-construction, rather than the individual. He and Young (1998) also strongly advocated an understanding of interactional competence that would encompass both co-construction and the inherently local nature of the participants' knowledge and interactive skills: "interactional competence is not an attribute of an individual participant, and thus we cannot say that an individual is interactionally competent" (p.7). Given this understanding, it is unsurprising that the possibility of evaluating the joint performances of candidates engaged in paired and group tests was raised (Swain in an interview with Fox, 2004). The possibility of shared scores has profound implications for defining and operationalising the construct of interactional competence in speaking tests, as Fulcher (2003: 46) acknowledges: "If talk in second language speaking tests is co-constructed…. we have to ask many questions, such as how

scores can be given to an individual test-taker rather than pairs of test-takers in a paired test format".

Another manifestation of the growing awareness of the inherent co-construction in speaking tests was the foregrounding of concern regarding the extent to which the discourse elicited through speaking tests that purported to be conversational in nature was actually so. This was questioned by van Lier (1989), He and Young (1998) and Johnson (2000, 2001). The asymmetric nature of these interviewer-led interactions led to serious concerns regarding the validity of inferences that could be made about a candidate's interactional competence on the basis of performance on this task. Essentially, if candidates were engaged in a form of "non-conversation", as Johnson (2001) maintained, this severely compromised what could be convincingly inferred about their ability to participate in a conversation in real life.

LANGUAGE TESTING INTERVIEWS: FOCUS ON INTERACTION

Language testing researchers have closely examined the validity of traditional interviews in relation to two high stakes tests of oral proficiency: the Oral Proficiency Interview (OPI) and the International English Language Testing System (IELTS). Studies on language testing interviews have explored the extent and type of interviewer accommodation to the interlocutor (including Brown, 2003, 2005; Brown & Hill, 1998; Lazaraton, 1996a; Ross, 1992, 1996; Ross & Berwick, 1992; Young & He, 1998;), cross-cultural pragmatics (Berwick & Ross, 1996) test-taker characteristics (O'Loughlin, 2000; O'Sullivan, 2000) and the construct validity of the "conversational" interview (Johnson, 2000, 2001; Lazaraton, 1992; Perret, 1990; van Lier, 1989;). A clear indication of the emerging focus on interactional competence in the language testing field was the publication of an influential collection of fourteen papers addressing validity issues related to co-construction in language testing interviews, edited by Young and He (1998). From these studies the emergence of two issues of crucial importance to the field emerged: the role and impact

of the trained interlocutor in the co-construction of discourse that was traditionally seen as a manifestation of the candidate's proficiency alone, and the validity of inferences that could be made about a candidate's ability to converse from performance on the relatively asymmetric, unnatural interaction that the language testing interview was shown to be.

The interlocutor effect

The impact of the interlocutor on the interaction in a language testing interview and the subsequent impact on rating has been the focus of a growing body of research into the IELTS speaking test by Brown and Hill (1998) and Brown (2000, 2003, 2005). After identifying interviewer difficulty through a multi-faceted Rasch analysis of score data, Brown (2005) was able to convincingly demonstrate that when candidates interacted with particular interviewers, they were more likely to be awarded higher or lower scores. In order to explore this phenomenon further, she selected four pairs of interviews for a closer analysis. These four pairs entailed the same candidate being interviewed by an interviewer previously identified as being "easy" or "difficult", in terms of the scores that candidates received when they were interviewing. Reactions of the raters to the interaction as it unfolded were captured by the use of stimulated verbal recall, which had been successfully used by DiPardo (1994) in the assessment of writing.

Using stimulated verbal recalls from raters, Brown (2005: 254) was able to identify "unhelpful" interviewer behaviours. Raters noted these features of interviewer behaviour when an unsatisfactory performance from the candidate could not be wholly attributed to the candidate's perceived level of spoken proficiency. This appeared to create a dissonance that prompted raters to notice aspects of the interviewer's management of the interaction that had impacted negatively on the candidate's opportunity to display their spoken proficiency. Interviewer behaviours noted by the raters included the extent to which interviewers asked questions that were either too difficult or too easy, and the way in which they developed

the topic: if it was too laboured this could result in a candidate having nothing more to say, but if topics were switched too rapidly a candidate might not be given the chance to elaborate their ideas. The use of closed questions and the adoption of a condescending or uninterested tone were other interviewer behaviours that raters perceived in a negative way.

In addition to identifying features that characterized "unhelpful" interviewer behaviours, Brown's study (2005) also identified features that characterized "helpful" interviewer behaviour. Her findings support those of Morton, Wigglesworth and Williams (1997), who were able to identify features of the ten "best" and "worst" interlocutors in their study. They characterized "good" interviewers as those who had "an encouraging, relaxed style that was responsive to the needs of the candidates", while poor interviewers "were less successful in modifying prompts where a candidate failed to understand the task or a word…or where a candidate gave an irrelevant response" (p.183). One of the important implications of these findings is that the previously held position that a trained interviewer was somehow a neutral factor in the co-constructed performance was no longer possible to uncritically accept.

Raters compensating for aspects of interviewer behaviour

Brown (2005) also explored the extent to which raters compensated for features of interviewer behaviour, and found that while raters did compensate for perceived unhelpfulness of the interviewer, "the notion of compensation appears to be particularly important in cases where the candidate is viewed as sitting between two levels on the scale. In these cases, raters may refer to their perception of the interviewer to justify awarding a higher rating than the performance would appear to warrant" (p.256). The phenomenon of raters compensating for perceived difficulties of the interactional style of their interlocutor had earlier been explored by McNamara and Lumley (1997), in a study on rater reactions to audiotaped speaking components of the Occupational English Test (OET) for health professionals. Raters were asked to not only rate the performance of

candidates, but also answer questions on the rapport that they perceived had been established between the trained "native speaker" interlocutor and the candidate and the competence of the interlocutor in carrying out their role. They found that while raters were not always in agreement regarding the perceived competence of the interlocutor, "perceptions of problems with interlocutor competence led to higher ratings" (p.152).

These findings had implications for interviewer and rater training, and also raised issues of fairness to candidates, particularly those who had the misfortune to be interviewed by an interlocutor who adopted the features of an "unhelpful" style. As Brown (2005) notes, the format of the IELTS speaking test on which she conducted her research has now changed, with the introduction of interlocutor frames to prescribe interviewer behaviour through set questions. However, she concludes that "it is not appropriate to assume that the variation that occurs in oral interviews interaction is *not* relevant to the construct "(p.262).

From the studies on the impact of the interlocutor in language testing interviews surveyed in this section, it is clear that the interlocutor, while trained and proficient in the language being assessed, can no longer be considered to have an "invisible" or neutral role in an interview which is taken to be a manifestation of the candidate's ability alone. Through problematising this issue, and using primarily qualitative research methodology to prove the impact of the interlocutor on both the discourse elicited from candidates and rater perceptions of proficiency, these studies have raised the question of whether this variability is an aspect that is irrelevant to the construct being assessed, and thus should be minimised through the use of interlocutor frames, or whether interlocutor variation is relevant to the construct. Those who advocate the former approach include Lazaraton (1996b), who detailed the introduction of interlocutor frames to CASE, and Morton, Wigglesworth and Williams (1997), who recommended from their study that as the most variation amongst interlocutor behaviour had occurred during the role-play segment of the

speaking test, this section should be more structured. However, the introduction of interlocutor frames fundamentally changes the nature of the interaction, and it is questionable as to whether inferences could be made regarding a candidate's capacity to interact in anything other than the "non-conversation" of a language testing interview. These decisions, as McNamara and Lumley (1997) point out, manifest a view of the construct that is being assessed: "it is important that the extent of these other variables be understood, both for theoretical reasons as part of our ongoing attempt to conceptualize the nature of performance assessment adequately, and for practical reasons in ensuring fairness to candidates" (p.145).

GROUP SPEAKING TESTS

If the ability of candidates to interact with others in more potentially symmetrical interactions than a language testing interview allows is considered a priority in task design, then it would seem that the group oral test format has the potential to maximise opportunities for this. Other advantages, including practicality, as evidenced by the need for fewer resources in terms of raters, time and rooms, make the group oral an attractive option to administrators. In addition, research has found that generally students feel positive about the group oral, with Fulcher (1996: 31) reporting that "engaging in a group discussion with a partner gave the students more confidence to speak and say what they wanted, rather than having to respond to an examiner". Positive reaction to the group oral from the perspectives of both teachers and candidates was reported by Hilsdon (1991), although she also raised concerns regarding reliability in scoring and the way in which the high-stakes nature of the test made a genuine group discussion hard to achieve.

Fulcher (1996: 38) also expressed concern regarding the lack of empirical evidence as to the nature of interaction that was occurring in group oral tests, and noted that this would involve analysis of candidate discourse. Three recent validation studies of group oral tests (He & Dai, 2006; Nakastuhara, 2010; Van Moere, 2010) have

explored candidate interaction through discourse analysis that addresses the earlier research agenda proposed by Fulcher.

He and Dai (2006) explored interaction in a group oral in the context of the College English Test –Spoken English Test (CET-SET) used in China, through an analysis of the segment in this test involving group discussion on a given topic. The format includes an interlocutor, examiner, and three to four candidates. Commenting on the lack of validation that had been carried out since the implementation of the test in 1999, they noted that: "CET-SET designers hold the view that it is a direct assessment of the candidate's ability in interactional competence in that speaking is a productive skill and its outcome can be directly observed" (p.377). In order to explore this claim, they transcribed and coded a 170,000 word corpus of CET-SET group oral performances, which enabled them to compare candidate discourse with the eight Interactional Language Functions (ILF) that the group discussion task was designed to elicit. He and Dai (2006: 385) found that *disagreeing* was the most frequently elicited ILF, accounting for almost half of the coded ILFs in the data set, with *asking for opinions or information* accounting for twenty four percent of the coded ILFs. The other six ILFs – *challenging, supporting, modifying, persuading, developing* and *negotiating meaning* – thus accounted for a very low percentage each.

He and Dai (2006) attributed the low occurrence of these ILFs to the candidates' framing of the task as an assessment event, rather than a real discussion, thus supporting Hilsdon's (1991) anecdotal evidence. The difficulty of trying to achieve a high degree of authenticity in assessment tasks is raised by Spence-Brown (2001), who concludes that "the fact that a task is used for assessment makes it unlikely that participants will engage with it in the same way that they would if they were not being assessed, no matter how much the assessment task resembles a real-world task in other aspects" (p.479). Thus, despite the intention of the group oral for candidates to interact in a more genuine way with each other than would be possible in a traditional language testing interview, He and Dai found that the candidates actually "consider the examiners, rather

than the other candidates in the group, to be their target audience" (p.389). This results in candidates avoiding negotiation of meaning and more complex functions such as challenging, while simply stating their own opinion in long turns, with He and Dai (2006) concluding that "it seems many candidates interpret contribution in terms of quantity rather quality" (p.391), which echoes Douglas' (1994) concerns regarding quality and quantity in speaking test performance. He and Dai also note that some group discussions actually resemble a series of short monologues, which is characteristic of a parallel pattern of interaction, identified by Galaczi (2004) in paired speaking tests.

Candidate and task variables

More recently, studies by Nakatsuhara (2010) and Van Moere (2010) have explored the role of candidate and task variables on interaction in group oral tests. While Nakatsuhara explored the interplay of test-taker variables, task types and group size on group dynamics and interactional patterns, the focus of Van Moere's study was the impact of different tasks on the elicitation of language functions. Nakatsuhara's study included 269 Japanese high school students, who were tested in groups on three speaking tasks: information gap, ranking, and free discussion. In addition to the task variable, candidates were categorized according to extroversion and oral proficiency levels. Extroversion levels were measured through the use of a Japanese version of the Eysenck Personality Questionnaire. The final variable was group size: candidates were placed in groups of either three or four. Through a complex research design where mixed methods were effectively utilized, Nakatsuhara found that while proficiency levels were influential in all tasks, extroversion-level variables had a greater impact in more open tasks, including the ranking and free-discussion tasks. Using CA to examine the turn by turn interactions, Nakatsuhara identified the features of the more closed information gap task including the compulsory information exchange and the information ordering of the task forcing the interactional roles and sequencing that minimized the impact of

extroversion levels. To add to the complexity of the findings, it appears that extroversion level variables had more impact on groups of four than groups of three, while proficiency level variables, while salient to both group sizes, had a larger impact in groups of three than in groups of four. Nakatsuhara concluded that the impact of task type, test-taker characteristics and group size should be taken into account by both test developers and researchers. She recommends that a group size of three is more suitable for oral tests, as interactional patterns became noticeably more artificial and introverted test-takers may contribute little in larger groups.

Van Moere (2010) explored the impact of task variables on candidate discourse and ratings, in the context of group orals used to assess speaking in the Kanda English Proficiency Test (KEPT). He compared candidate discourse elicited through three tasks: a group discussion task, a consensus task and a picture difference task through an analysis of word/turn count ILFs and scores. Van Moere found that each task elicited a different frequency and range of ILFs, with the picture task frequently eliciting negotiation of meaning, while the consensus task elicited the widest range of ILFs. The importance of a specific goal for interaction was noted by Van Moere, as discussions that were unfocussed or not goal-oriented did not consistently elicit authentic conversations. He also found evidence of parallel patterns of interaction with candidates in some groups orienting to short monologues with a token "how about you" to nominate the next speaker, thus supporting findings from He and Dai (2006).

Emerging from these studies is a deeper understanding of the complex interplay of task and candidate variables in group oral tests. The question remains as to which of these variables are considered salient to the construct, and which are irrelevant. If different tasks elicit a differing variety and frequency of ILFs, there would seem to be an argument for including a range of tasks in a group speaking test. The impact of task and candidate variables needs to be studied in a range of testing contexts, and further research on the stability of a candidate's level of extroversion would also be useful. Noticeably

absent from the existing studies into group orals is the rater, and features of the performance that influence rater decisions.

PAIRED SPEAKING TESTS

Paired speaking tasks are currently used in high-stakes speaking test contexts, including the University of Cambridge ESOL examinations First Certificate in English (FCE) and Certificate in Advanced English (CAE). While the paired format would seem to offer the potential for candidates to interact in a more symmetrical way than the traditional language testing interview allowed, concern regarding the paucity of validation studies into speaking tests involving candidate-to-candidate communication has been raised by Foot (1999a, 1999b) and Swain (2001). Foot was particularly concerned with the potential mismatching of candidates in terms of spoken proficiency and the prospect of mutual incomprehensibility resulting from pronunciation errors or strong accents.

Two of the earliest studies into paired candidate speaking tests encompassed the impact of the interlocutor in terms of the respective proficiency levels of the paired candidates (Iwashita, 1998) and the impact of the familiarity of candidates (Ikeda, 1998). These two studies helped to set a research agenda for later researchers, and issues raised in them continue to be explored today. Iwashita (1998) compared the impact on candidates' scores and discourse when paired with an interlocutor of a similar and different proficiency level. The participants were twenty adult learners of Japanese. She found that although the proficiency of the interlocutor did impact on the quantity of discourse elicited through the task, it did not seem to significantly change scores given to candidates. In addition, test-taker feedback indicated that "candidates prefer the NNS-NNS interaction mode to the NS-NNS mode as they find it less threatening" (p.52). Candidate preference for the paired candidate interaction was also found by Taylor (2001) and May (2000).

Ikeda (1998) explored the paired candidate interaction from a Vygotskian perspective. Through a study of five "paired learner

interviews", involving teenage Japanese students of English, he found that this testing task offered the candidates opportunities not only to negotiate meaning, but also to "take initiative to learn new knowledge and incorporate it into their respective private worlds" (p.71). Ikeda allowed candidates to select their interlocutor, and cautioned against the "risk of pairing linguistically compatible learners who may be incompatible personality-wise" (p.93).

The focus of research into paired candidate speaking tests was later extended to a comparison of test-taker feedback comparing attitudes to interviews and paired candidate speaking tests (May, 2000) and a comparison of speaking functions elicited through interviews and paired candidate speaking tests (Taylor, 2001).

In an exploratory study with 32 EAP students from China, May (2000) compared test-taker reactions to the use of a traditional oral proficiency interview and a paired candidate speaking test. She found that candidates not only preferred the paired candidate speaking test, but were aware of the power differential inherent in the oral proficiency interview, and viewed the opportunity for a genuine exchange of views and the exposure to and creation of new "knowledge" as advantages of the paired candidate speaking test (p.17), thus supporting Ikeda's (1998) findings. Interestingly, the candidates themselves were not concerned about the impact of being paired with a partner of a differing language proficiency, with one candidate positioning the performance as being inherently the product of an individual: "the interaction is bilateral, but the ideas presented and the way of doing the task are still unilateral. So the partner would not affect the results" (p.17).

Responding to Foot's (1999a) criticism on behalf of UCLES, Taylor (2001) reported the results from two internal studies which had been undertaken in order to compare paired and one-to-one speaking test formats. The paired speaking test format was shown to elicit more *informational functions* and *managing interaction functions* than the one-to-one interview. In addition, whereas *informational functions* made up approximately 80% of the candidates' discourse in

the one-to-one interview, they only accounted for 55% of the candidates' discourse in the paired speaking test format (p.16). From this Taylor concluded that paired speaking tests have the potential to be more symmetrical and genuinely interactive than traditional one-to-one interviews. Also cited in this report were findings that of a possible 30 speaking functions, 26 were evident in the paired speaker tests, while only 14 were found in the one-to-one interviews. However, these results were gained from a very small sample: only three paired speaking tests and three one-to-one tests, which underscored the need for further research to be carried out on paired speaking tests.

Egyud and Glover (2001) also responded to Foot's criticism with a spirited defence of oral testing in pairs within a secondary school context. Reporting on test-taker feedback from teenage Hungarian learners of English, they found that candidates both liked paired tests and felt they were less stressful than traditional interviews. Through examining the discourse produced by candidates in a paired direction task with the same task undertaken with an "expert" interlocutor, they found that the candidate performed "better" in the paired task, though this is based on a discourse analysis of just one paired speaking test. Concern over the lack of published validation studies carried out into tests requiring candidates to interact with each other was expressed by Swain (2001), who strongly recommended that candidate discourse be examined. A number of studies on paired speaking tests were subsequently published, with the work of Galaczi (2004), Nakatsuhara (2004), Brooks (2003, 2004) and Lu (2003a, 2003b) explicitly examining discourse in order to explore the interactional patterns elicited.

Examining candidate discourse

Using CA to thoroughly explore turn by turn sequences of interaction in the Cambridge FCE, Galaczi (2004) convincingly categorised the dyadic patterns of discourse co-construction into three main types: collaborative interaction, parallel interaction,

asymmetric interaction. If candidates oriented toward more than one of these interactional patterns, it was termed a 'blended' interaction. The basis for Galaczi's (2004) categorisation of the patterns of discourse co-construction lies in the extent of mutuality and equality evident in each paired candidate segment of the FCE. Collaborative interactional patterns were characterized by both partners taking the opportunity to introduce topics, and develop their partner's topic, thus exhibiting high equality and high mutuality. In contrast, parallel interactional patterns were characterized by "solo vs. solo" (p.254) performances from the candidates. While candidates were able to initiate topics, they were unlikely to respond to their partner's topic initiation by developing it, and Galaczi (2004: 254) notes that the speakers were "much more concerned with developing their own contributions instead of engaging with each other's contributions". Asymmetric interactions involved a dominant and a passive speaker, with the dominant speaker contributing "more to the task while the passive speaker oriented to a more reactive role" indicating low equality. In addition to these three patterns of interaction, Galaczi also documented a fourth pattern, which she terms "blended". In blended interactions, where features associated with two patterns of interaction were manifested, "typically a dyad would alternate from one pattern to another" (p.257). In a data set comprised 30 paired candidate performances from the FCE, she found that while the majority of the test-taker dyads "oriented either to a collaborative (30%), parallel (30%), or blended (30%) pattern of interaction….. asymmetric dyads ….. comprised 10% of the dataset" (p.112). As Galaczi notes, it is the asymmetric dyads that "are potentially the most problematic from an assessment perspective" (p.261).

Galaczi drew on Storch's (2001: 113) model of dyadic interactions in a paired learner task in a classroom context, where four interactional patterns were identified: collaborative, which was characterized by moderate to high equality and moderate to high mutuality; dominant/dominant, characterized by moderate to high equality but moderate to low mutuality; dominant/passive,

characterized by moderate to low equality and moderate to low mutuality; expert/novice, characterized by moderate to low equality but moderate to high mutuality. In addition to stressing that the interactional patterns occurred along a continuum – hence the moderate to high description, for example – Storch cautioned that "categorisation by its very nature is imprecise" (p.115).

In developing Storch's (2001) model and adapting it to the context of paired speaking tests, Galaczi (2004) points out that a sub-category of the asymmetric interactions, which she categorises as "low dominance" where one partner is forced into the role of dominance by the passivity of the other partner, exhibits similar features to Storch's (2001) definition of the "expert-novice" pattern of interaction, where the dominant partner plays a supportive role in that they encourage the more passive partner in ways similar to that of a teacher.

Figure 1 illustrates Galaczi's summary of the features which characterise collaborative, parallel and asymmetric interactions in a paired speaking test, including topic "life", mutuality and equality.

The identification of discourse features accompanying higher and lower scores awarded by raters for "Interactive Communication" was another important finding by Galaczi (2004). While Galaczi (2004: 264) speculates that "the conversational management ability of L2 learners has a higher and lower level, with collaborative dyadic interaction being the higher-level skill, and parallel dyadic interaction being the lower-level skill", this phenomenon needs further research, as it has implications for the pairing of candidates and rating scale development. It could also be interpreted as positioning collaborative interaction as the "gold standard" of communication, regardless of context or communicative purpose. This is questionable, as in real life we may be required to achieve our communicative purpose while interacting with someone who is more powerful in a particular hierarchy, or who is attempting to dominate an interaction.

FIGURE 1

Galaczi's (2004: 184) Summary of the characteristics of the collaborative, parallel and asymmetric patterns of interaction

Interactional characteristics	Collaborative interaction	Parallel interaction	Asymmetric interaction	
			Dominant speaker	Passive speaker
Mutuality	High	Low	Low/High	
Equality	High	High	Low	
Topic "life'	Long	Short	Moderate	
Structure of proto-typical topic development sequences	A: Topic initiation + Topic building	A: Topic initiation + Topic building	A: Topic initiation + Topic building	
	↓	↓	↓	
	B: Topic extension	B: Minimal acknowledement + Topic initiation	Minimal acknowledgement	
	↓	↓	↓	
	A: Topic extension + Topic initiation	A: Minimal acknowledgement + Topic initiation	A: Topic extension	

While reflecting on lower levels of inter-rater agreement when rating dyads orienting to asymmetric interactional patterns, Galaczi (2004, p.262) suggests the "strong possibility that factors other than language proficiency" may be responsible for the resulting patterns of interaction, including culture and personality. Her concern echoes that of Fulcher (2003) and of Taylor (1996), who provides one of the few published perspectives of an experienced rater of CAE paired candidate speaking tests, and voices his doubts over certain aspects of the test. Through a description of one particular test situation, he illustrated the difficulties of assessing "Interactive Conversation", and questions the extent to which cultural factors could lead to difficulties. His call for CA to be used for a deeper understanding of exactly what constitutes effective interaction has been echoed by Swain (2001) and has begun to be addressed by research reported on in this section.

Galaczi's most recent study (2010) explored the features of interactional competence that were salient at different oral proficiency levels in paired speaking tests, and the extent to which features of interactional competence can be meaningfully operationalised in assessment scales. Using CA to explore the features of interaction in thirty-two paired discussion tasks from a Cambridge ESOL Main Suite exam, Galaczi identified aspects of topic development strategies, topic life, listener support and turn-taking at four levels. These detailed descriptions were based on observable interactional behaviours, and have the potential to both enhance our understanding of the construct of interactional competence in paired speaking tests, and inform the development of more meaningful rating scales.

Brooks (2003, 2004, 2009) reported a comparison between candidate performance in paired candidate speaking tests and an "individual" format, which is actually a speaking test where candidates interact with a teacher. Brooks' (2004) claim that paired candidate speaking tests and teacher-led interactions are "interactionally different" points to a key issue for the field to

address. While Taylor (2001) reported that paired speaking tests enabled candidates to demonstrate a wider range of interactional skills than interviewer-led tests, she appeared to view these as being on a continuum, rather than representing completely different constructs. If paired speaking tests and teacher-led interactions do indeed tap into different constructs, then the case for including both test formats into speaking tests would appear to be strengthened. The issue is to define and substantiate those constructs, which remains for the field to address in a convincing manner.

Focus on candidate variables: proficiency, gender and first language

The impact of the pairing of candidates, which had been raised as a concern by Foot in 1999, was explored by O'Sullivan (2002), Lu (2003), Csepes (2002), Nakatsuhara (2004), Norton (2005) and Davis (2009). Learner acquaintanceship, which Ikeda has explored in 1998, was the focus of O'Sullivan's study (2002), in which a group of 32 adult Japanese learners of English participated in two paired interactions, one of which was with a friend, the other with an unfamiliar interlocutor. Although he suggests that there may be evidence of "an acquaintanceship effect, with subjects achieving higher scores when working with a friend" (p.277), he found that there appeared to be no impact on the linguistic complexity of discourse elicited.

Lu (2003) explored the impact of candidates' first language on their "discoursal performance" in a paired candidate speaking test, and speculated on the impact of a shared first language among paired candidates. Lu concluded that "the most recurrent discourse features produced by test-takers in paired-format OPTs seem to be influenced by their first languages" but that "in terms of overall discoursal performance test-takers may not be disadvantaged if paired with someone who shares the same or a different first language" (2003, conference handout). This has important implications, as many speaking tests are carried out in countries

where candidates will be interacting with another candidate sharing the same first language.

The impact of pairing candidates of differing language proficiency, which had earlier been raised by Foot (1999a), was the focus of several recent studies on paired speaking tests. Csepes (2002) carried out a primarily quantitative study of partner effects on oral test scores in the context of Hungarian secondary schools. She paired candidates with a partner of lower, similar and higher language proficiency, and then compared scores given to each candidate on the three occasions. She concluded that scores given by the raters "suggest that their perception of core students' proficiency was neither positively nor negatively influenced" by the language proficiency of the candidate they were paired with. Nakatsuhara (2004) also explored the effect of pairing candidates of the same (SPL) and different language proficiency (DPL) levels in the section of the CAE tests in which candidates collaborate in a problem solving discussion. Through an analysis of discourse, she found that "the pairing of the students with different language levels may not be as problematic as expected"(p.57). This conclusion was reiterated by Davis (2009), who placed Chinese undergraduate candidates in two paired speaking tests: one with a partner of a similar level, the other with a partner of either a higher or a lower proficiency level. He found that despite differences in the quantity of language, "the proficiency level of an examinee's partner in a paired oral test had little influence on scores" (p.388).

While the findings of Davis (2009), Norton (2005), Nakatsuhara (2004) and Csepes (2002) indicate that partnering candidates of differing levels of proficiency in paired candidate speaking tests did not have a substantial impact on opportunities for candidates to display their interactional competence, Brown's (2003, 2005) research on interviewer-led speaking tests found that the interlocutor's interactional style had a significant impact on the rater's judgments of a candidate's proficiency. These findings appear to be contradictory, in the sense that the impact of the interlocutor in paired speaking tests appears to be minimal, while in interviewer-

led tests, raters were cognizant of the ways in which "unhelpful" interviewer behaviours resulted in limited opportunities for candidates to display their spoken proficiency. The different results may be explained by the methodology employed by Brown, as she not only studied candidate discourse and scores awarded by raters, but also the raters' decision making in terms of reasons for giving those scores. If we only consider candidate discourse and scores given, we can only infer the reasons that raters gave scores.

Focus on the rater

Recent studies that went beyond scores to explore how raters construed and operationalised interactional competence have been carried out by Ducasse and Brown (2009) and May (2007, 2009). In a study of raters commenting on performances of paired beginner learners of Spanish, Ducasse and Brown found that the three main categories of interactional features salient to raters were non-verbal interpersonal communication, interactive listening and interactional management. Their findings on the importance of interactive listening, which they categorised into "supportive listening" and "comprehension" reinforce the need for research that encompasses both listening and speaking dimensions in speaking assessment. Their study has the potential to inform the development of rating scales that reflect the complexity of both speaking and listening constructs in interaction.

May (2007) categorised nineteen features of interactional competence that raters found salient when assessing a paired EAP discussion task. These features included assertiveness through communication, conversation management and body language. Interactive listening was also salient to the raters as an essential element of both comprehending and responding to a partner. When raters were faced with asymmetric interactions, with one partner clearly dominating, they found it difficult to separate the impact of one candidate on the other and award separate scores for interactional competence (May, 2009). In these situations the raters would either compensate a candidate based on what they might

have been able to achieve with a different partner, or penalize one or both candidates for their role in co-constructing the asymmetric interaction.

This section has demonstrated that the research agenda with regard to paired speaking tests has already incorporated studies into candidate variables, candidate discourse and rater operationalisation of interactional competence. Larger scale studies in a variety of assessment contexts would help to clarify and continue to address key issues including the impact of interlocutors of differing proficiencies, the separability of the candidates, the development of rating scales that more fully reflect the complexity of interaction, and the role of interactive listening in paired speaking tests.

QUESTIONS FOR FUTURE RESEARCH

The previous sections of this review have outlined key studies relating to interactional competence that have informed our understanding of this complex and multifaceted construct. A hallmark of many of the studies is the utilization of qualitative research methodologies, and in particular, the analysis of discourse, which was strongly advocated by Shohamy (1998), He and Young (1998) and Swain (2001). Surveys of discourse and assessment by McNamara, Hill and May (2002) and Young (2002) have highlighted the extent to which Conversation Analysis (CA) is now being used by language testing researchers, and the insights which can be gained from its use, particularly when interaction is the focus of research. McNamara and Roever (2006: 46) describe the position of CA as one in which "the interlocutor is implicated in each move by the candidate; that is, the performance is a dance in which it makes no sense to isolate the contributions of the individual dance partners". It is thus ideally positioned as a means to explore co-construction.

The combination of CA and rater studies has the potential to inform us of not only the observable features of an interaction, but also the aspects of the performance that are salient to raters. While

several rater studies have been carried out in the context of paired orals, there is little yet known about the features of interactional competence that are salient to raters of group oral tests. Studies combining a focus on the rater and candidate discourse would help in the development of rating scales that could more meaningfully incorporate key features of interactional competence, including interactive listening. The perspective of the candidates would also be valuable to ascertain, and in particular the way that candidates frame the interaction as it unfolds.

In terms of responding to the difficulties inherent in operationalising key aspects of interactional competence, three approaches seem to be apparent in current high-stakes speaking tests. The first approach is to standardize the contributions of the interlocutor in a language testing interview through the use of interlocutor frames, as is the case in the IELTS speaking test. While the use of interlocutor frames can help to guard against candidate performance being adversely affected by an interviewer's style, it has implications for the inferences that can be made on the basis of an interaction that has largely been pre-scripted, in terms of the interviewer's contribution. The second approach is to reflect the complexity of the construct by providing candidates with a variety of interlocutors and tasks, as is done in the FCE. Candidates in this test have the opportunity to interact with both a trained interlocutor using an interlocutor frame and with another candidate in a discussion task as they proceed through a series of interactions, each with a different focus. The third approach, clearly manifested through semi-direct computer-based tests of speaking including the TOEFL iBT, is to limit the candidate responses to short monologues, thus focussing on a 'solo' speaking performance.

Ultimately, the test purpose and context must be considered when decisions are made to either embrace or deliberately omit certain features of interactional competence. In the case of high-stakes testing contexts the focus will not be on a "faithful account of the interaction" in its richness and complexity, but, as McNamara and Roever (2006: 51) point out, "a score about individual candidates

that can then be fed into institutional decision making procedures". How to reconcile this need for a score that reflects an individual candidate's speaking proficiency with the complexity and localized nature of interactional competence is a question that continues to challenge the field.

THE AUTHOR

Lyn May is a lecturer in TESOL at the Queensland University of Technology. Her research in language assessment focuses on the validation of speaking tests, interactional competence and the oracy demands of tertiary study. Lyn has published on assessment in *Language Testing*, *Annual Review of Applied Linguistics*, and *Melbourne Papers in Language Testing*.

REFERENCES

Berwick, R. & Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp.34-54). Studies in Language Testing 3. Cambridge: Cambridge University Press.

Brooks, L. (2003). *An investigation of the interactions in paired oral proficiency testing*. Paper presented at the 25th Language Testing Research Colloquium, Reading, United Kingdom.

Brooks, L. (2004). *Insights into the construct(s): Paired oral proficiency testing*. Paper presented at the 26th Language Testing Research Colloquium, Temecula, United States.

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

Brown, A. (2000). An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (Ed.), *IELTS*

Research Reports, 3 (pp.49-84). IELTS Australia and the British Council.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews.* Frankfurt: Peter Lang.

Brown, A. & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Wood (Ed.), *IELTS Research Reports*, 1 (pp.1-19). IELTS Australia and the British Council.

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

Csepes, I. (2002). *Measuring oral proficiency through paired-task performance.* PhD dissertation. Budapest.

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

DiPardo, A. (1994). Stimulated recall in research on writing: An antidote to "I don't know, it was fine". In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp.163-181). Thousand Oaks, CA:Sage.

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

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

Egyud, G. & Glover, P. (2001). Oral Testing in pairs: A secondary school perspective. *ELT Journal*, 55(1), 70-76.

Foot, M.C. (1999a). Relaxing in pairs. *ELT Journal*, 53(1), 36-41.

Foot, M.C. (1999b). Reply to Saville and Hargreaves. *ELT Journal*, 53(1), 52-53.

Fox, J. (2004). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly*, 1(4), 235-251.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13, 23-49.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson.

Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: the case of the First Certificate in English*. PhD dissertation: Columbia University.

Galaczi, E. (2010). *Interactional competence across proficiency levels.* Paper presented at the Language Testing Research Colloquium, April 2010, Cambridge.

Hall, J.K. (1995). (Re)creating our worlds with words: a sociohistorical perspective of face-to-face interaction. *Applied Linguistics*, 16(2), 206-232.

He, L. & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, 23(3), 370-401.

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.1-4). Amsterdam: John Benjamins.

Hilsdon, J. (1991). The group oral exam: advantages and limitations. In J.C. Alderson & B. North (Eds), *Language testing in the 1990s: the Communicative legacy* (pp.189-197). London: Modern English Publications and the British Council.

Ikeda, K. (1998). The paired learner interview: A preliminary investigation applying Vygotskian insights. *Language, Culture and Curriculum*, 11, 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(3), 171-183.

Johnson, M. (2000). Interaction in the oral proficiency interview: problems of validity. *Pragmatics*, 10(2), 215-231.

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the Oral Proficiency Interview*. 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. (1992). The structural organization of a language interview: a conversation analytic perspective. *System*, 20(3), 373-386.

Lazaraton, A. (1996a). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, 13, 151-172.

Lazaraton, A. (1996b). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp.18-33). Studies in Language Testing 3. Cambridge: Cambridge University Press.

Lu, Y. (2003). *Test-takers' first languages and their discoursal performance in paired-format OPT*. Paper presented at the 25th Language Testing Research Colloquium, Reading, United Kingdom.

May, L. (2000). Assessment of oral proficiency in EAP programs: A case for pair interaction. *Language and Communication Review*, 9(1), 13-19.

May, L. (2007). *Interaction in a paired speaking test: The rater's perspective*. PhD thesis. University of Melbourne.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397-422.

McNamara, T.F. (1996). *Measuring second language performance.* Harlow: Addison Wesley Longman.

McNamara, T.F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, 18, 444-446.

McNamara, T.F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, 14(2), 140-156.

McNamara, T. F., Hill, K. & May, L. (2002). Discourse and Assessment. *Annual Review of Applied Linguistics*, 22, 221-242.

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

Morton, J., Wigglesworth, G. & Williams, D. (1997). Approaches to the evaluation of interviewer behaviour in oral tests. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery*. NCELTR Research Series 9, 175-195.

Nakatsuhara, F. (2004). *An Investigation into Conversational Styles in Paired Speaking Tests.* MA dissertation: University of Essex.

Nakatsuhara, F. (2010). *Interactional competence measured in group oral tests: how do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups?* Paper presented at the Language Testing Research Colloquium, April, 2010, Cambridge.

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

O'Loughlin, K. (2000). The impact of gender in the IELTS oral interview. In R. Tulloch (Ed.), *IELTS Research Reports*, 3 (pp.1-28). IELTS Australia and the British Council.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, 28, 373-386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295

Perrett, G. (1990). The language testing interview: a reappraisal. In J. de Jong and D.K. Stevenson (Eds.), *Individualising the assessment of language abilities* (pp.225-238). Philadelphia: Multilingual Matters.

Ross, S. (1992). The discourse of accommodation in oral proficiency inteviews. *Studies in Second Language Acquisition*, 14, 159-176.

Ross, S. (1996). Formulae and inter-interviewer variation in oral proficiency interview discourse. *Prospect*, 11(3), 3-16.

Ross, S. & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14(2), 159-176.

Shohamy, E. (1998). How can language testing and SLA benefit from each other? The case of discourse. In L. Bachman & A. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp.156-176). Cambridge: Cambridge University Press.

Spence-Brown, R. (2001). The eye of the beholder: authenticity in an embedded assessment task. *Language Testing,* 18 (4), 463-481.

Storch, N. (2001). *An investigation into the nature of pair work in an ESL classroom and its effect on grammatical development*. PhD dissertation. The University of Melbourne.

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. (2001). The paired speaking test format: recent studies. *Research Notes*, 6, 15-17. Cambridge: University of Cambridge ESOL.

Taylor, R.E. (1996). Assessing oral communication skills- Reflections of an examiner. *World Englishes*, 15(1), 131-137.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly*, 23, 480-508.

Van Moere, A. (2010). *Group oral tests: What kinds of tasks and functions are optimal for eliciting and measuring interactional competence*? Paper presented at the Language Testing Research Colloquium, April, 2010, Cambridge.

Young, R. (2000). *Interactional Competence: Challenges for validity*. Paper presented at the Annual Meeting of the American Association of Applied Linguistics, March, 2000, Vancouver, Canada.

Young, R. (2002). Discourse approaches to oral language assessment. *Annual Review of Applied Linguistics*, 22, 243-62.

Young, R. & He, A.W. (Eds.) (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamins.