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

Co-constructed interaction in a paired speaking test: the rater's perspective

Lyn May
Queensland University of Technology

Abstract

The definition and operationalisation of interactional competence in speaking tests that entail co-construction of discourse is an area of language testing requiring further research. This article explores the reactions of four trained raters to paired candidates who oriented to asymmetric patterns of interaction in a discussion task. Through an analysis of candidate discourse combined with rater notes, stimulated verbal recalls, rater discussions and scores awarded for interactional effectiveness, the article examines the extent to which raters compensate or penalise candidates for their role in co-constructing asymmetric interactional patterns. The article argues that key features of the interaction are perceived by the raters as mutual achievements, and it further suggests that the awarding of shared scores for interactional competence is one way of acknowledging the inherently co-constructed nature of interaction in a paired speaking test.

I Introduction

The importance of defining and operationalising interactional competence in speaking tests was highlighted by Kramsch (1986, p. 368), who argued strongly for deeper understanding of a construct that she felt was overlooked by the proficiency movement: “by stressing behavioural functions and the lexical and grammatical forms of the language, the proficiency guidelines emphasize the static content structure, not the dynamic process of communication”. The potential disjunct between communicative language teaching pedagogy which encouraged classroom interaction in the form of group and pair work and speaking tests that focused on the candidate as an individual was highlighted by Kramsch.

The term that best encompasses one of the tenets of Kramsch’s (1986) concept of interactional competence is “co-construction”, defined by Jacoby and Ochs (1995, p. 171) as incorporating a “range of interactional processes, including collaboration, cooperation, and coordination”. If we accept the definition of speaking as “the use of oral language to interact directly and immediately with others” (Butler, Eignor, Jones, McNamara and Suomi, 2000, p. 2), it is apparent that the co-construction of discourse is central to successful interaction. While focusing on the “co” in co-construction, Jacoby and Ochs also noted that co-construction does not necessitate mutually supportive interaction, which has implications for all speaking tests, but particularly those with formats incorporating interaction between candidates, including paired and group orals. The issue of the shared responsibility for interactional patterns that interlocutors orient toward is raised by Jacoby and Ochs (1995, p. 177): “One of the important implications for taking the position that everything is co-constructed through interaction is that it follows that there is a distributed responsibility among interlocutors for the creation of sequential coherence, identities, meanings and events”.

In calling for the "intrinsically social nature of performance" to be recognized, rather than downplayed or factored out of speaking tests, McNamara (1996, 1997) addressed the issue of co-construction in speaking tests by problematising the assumption of communicative competence as residing in the individual, and argues that "a weakness of current models is that they focus too much on the individual rather than the individual in interaction" (1996, p. 85). He and Young (1998, p. 7) 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". As a way of acknowledging this phenomenon, Swain (2005, in an interview with Fox), suggests evaluating the joint performance, rather than the individual's contribution to it, which has profound implications for defining and operationalising the construct of interactional competence in paired speaking tests, as Fulcher (2003, p. 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".

In research on the impact of co-construction of the discourse between candidate and interviewer in the International English Language Testing System (IELTS) speaking test, Brown and Hill (1998) and Brown (2003, 2004) found that raters' perception of a candidate's oral proficiency, and hence the scores they awarded, were affected by the choice of interviewer. The introduction of interlocutor frames in IELTS to standardise interviewer contributions to the interaction could be interpreted as a response to the "construct irrelevant" impact of interlocutor variation. While attempting to eliminate, or at least minimize this aspect of the interaction may be considered important to enhancing reliability and fairness to candidates, it seems to be underpinned by a belief that performance in a speaking test can be assessed as a product of the candidate alone, which has implications for validity. As Weir (2005, p. 153) concludes: "the real problem is that an individual's performance is clearly affected by the way the discourse is co-constructed by the person they are interacting with. How to factor this into or out of assessment criteria is yet to be established in a satisfactory manner".

Brown (2004) also explored the extent to which IELTS raters compensated for features of interviewer behaviour, and found that raters did compensate for perceived unhelpfulness of the interviewer. This supports the findings of McNamara and Lumley (1997), who analysed 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" (McNamara & Lumley, 1997, p.152).

In contrast to the many published studies on the Oral Proficiency Interview (OPI) and IELTS, paired candidate speaking tests, which are also used in high stakes language assessment contexts, including several University of Cambridge English for Speakers of Other Languages (ESOL) examinations, have received relatively little attention in the language testing literature until recently. The earlier research into paired candidate speaking tests encompassed the impact of the interlocutor in terms of differing proficiency levels (Iwashita,1998), and candidate familiarity (Ikeda, 1998), test-taker feedback comparing candidate reactions to taking a paired speaking test and a traditional language testing interview (May, 2000) and a comparison of speaking functions elicited through interviews and paired candidate speaking tests (Taylor, 2001).

Swain (2001, p. 277) echoed Foot's (1999a, 1999b) concern over the lack of published research into speaking tests involving candidates interacting with each other: "given that small group testing occurs in even one high-stakes test, as well as its reasoned use, it is surprising that so little validation work has been carried out". Linking McNamara's (1997) concerns relating to the co-constructed discourse being regarded as the product of individual candidates only, Swain (2001, p. 297) recommended that candidate discourse be examined, so that a deeper understanding could be reached about exactly what was being elicited through pair and group test tasks, which could "provide test-developers with targets for measurement". A number of studies on paired speaking tests were subsequently published, with the work of Galaczi (2004), Nakatsuhara (2004), and Lu (2003) explicitly examining discourse in order to explore the interactional patterns elicited.

In the most thorough study on paired candidate discourse to date, Galaczi (2004) explored turn by turn sequences of paired candidates in the Cambridge First Certificate in English (FCE) through conversation analysis. Drawing on Storch's (2001) model of dyadic interactions, Galaczi identified four main patterns of discourse co-construction: collaborative, parallel, asymmetric, and blended interaction. The basis for this categorisation 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" (Storch, 2001, 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, p. 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" (Galaczi, 2004, p. 256) 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" (Galaczi, 2004, p. 257). In a data set comprised of 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". As Galaczi (2004, p. 112) notes, it is the asymmetric dyads that "are potentially the most problematic from an assessment perspective".

The impact of pairing candidates of similar and different speaking proficiency levels has been explored by Csepes (2002), Nakatsuhara (2004) and Norton (2005), with differing findings pointing to the need for further research. Csepes' (2002, p. 88) findings that the scores given by raters "suggest that their perceptions of core students' proficiency was neither positively nor negatively influenced by the fact that the level of proficiency of core students' partners showed considerable variation", are supported by Nakatsuhara (2004, p. 57), who concluded that whether students are paired with partners of similar or different proficiency levels, "they are likely to obtain rather identical opportunities to display their communicative abilities" and thus "the pairing of the students with different language levels may not be as problematic as expected". However Norton (2005, p. 291) suggests that "being paired with a candidate who has higher linguistic ability may be beneficial for lower level candidates who are able to incorporate some of their partner's expressions into their own speech".

From the research presented in this section, clearly emerging issues of concern with respect to paired speaking tests are the impact of one candidate on another in terms of scores awarded and discourse elicited. However, one important area remains unexplored: the rater's perspective, and with it, the question of the separability of the individual candidate's contribution and performance in a paired speaking test. In order to explore the rater's perspective, this article will focus on a paired speaking test which caused difficulty for raters as initially evidenced through the range of scores given to the candidates for interactional effectiveness.

II The study

1 Research question

The data presented in this article is designed to explore the question:

Are the contributions of individual candidates to the interactional patterns that emerge in paired candidate speaking tests considered separable by the raters?

2 Participants: Candidates

Twelve adult English for Academic Purposes (EAP) students from China participated in this study. The basis for selection of the candidates was that they formed two distinct groups: six candidates (three male and three female) were from an intermediate class, and six candidates (three male and three female) were from an advanced class. Their oral proficiency levels were confirmed by experienced EAP teachers who had taught them for the previous six months.

Participants: Raters

Four trained raters participated in the study. Raters 1, 3 and 4 were female, and Rater 2 was male. The system for rating paired candidate speaking tests in their university language centre involved two raters observing a paired candidate speaking test, making notes as they watched, and then coming to an individual rating decision. Immediately following this, the two raters are required to discuss the performance and their ratings, and come to a jointly negotiated final rating. Raters 1 and 2 were paired, as were Raters 3 and 4. Raters 2, 3 and 4 had all had several years experience rating paired candidate speaking tests at a university language centre, whereas Rater 1 was trained but had not previously rated. Raters 3 and 4 had rated together on several occasions prior to this. Raters 1 and 2 had never rated together.

Figure 1 Shared characteristics of raters

- speakers of English as a first language
- post-graduate TESOL qualifications
- five or more years of EAP teaching experience
- experienced EAP speaking teachers
- trained to rate paired speaking tests

3 Task

A structured discussion task, with a problem/issue and three proposals to address the problem was chosen for the paired speaking test. The candidates had previously been given readings related to the topics, and had also covered these topics in class discussions. The task sheet instructed candidates to discuss the value of each proposal, then indicate which of the proposals they think would be most effective, giving reasons for their choice, and finally, having considered the points raised in the discussion, come to an agreement, compromise or 'agree to disagree' with each other. This type of discussion task is used in a high stakes test of speaking at a university language centre in Australia. After being given the task sheets, the candidates had up to five minutes of planning time to prepare for the discussion, and twelve minutes to complete the discussion once they had begun.

4 Pairing of candidates

Candidates took two parallel forms of the discussion task (one on cloning, the other on genetically modified food), but on a different day and with a different partner. Thus each candidate had the opportunity to interact with a partner of the same level of spoken proficiency, and with a partner of a different level of spoken proficiency.

5 Rating Scales

An analytic rating scale, incorporating five categories- Fluency; Accuracy; Range; Effectiveness; and Overall was used by the raters. While the scale had been developed by experienced teachers and used for several years, it had not been systematically validated. All raters had been trained to use this scale. The rating scales provide a brief description of performance under each of the five categories for Bands 1-5.

The descriptors for "Effectiveness" include only three aspects: the extent to which the interlocutor's message was understood, the ability to respond to an interlocutor, and the use of communicative strategies. Thus the descriptor for a Band 4 reads:
Able to understand the interlocutor's message and is mostly able to respond accordingly. Is generally able to use effective communicative strategies.

6 Data collection

The 12 paired speaking tests (six involving candidates at similar levels, six involving candidates at different levels) were videoptaped. The data collection encompassed all aspects of the rater's usual rating process, which included a paired rater discussion at the end of each performance. In addition, raters produced a stimulated verbal recall after making an initial rating decision, and prior to the paired rater discussion. The collection and analysis of a wide range of data including the initial ratings, rater notes, summary comments, stimulated verbal recalls and paired rater discussions was designed to facilitate a multidimensional view of the rating process, and the features of a paired candidate speaking test that were salient to raters in their operationalisation of interactional effectiveness. In this way numerous sources of “evidence” about how raters think about interaction were encompassed, and the complexity is such that some of these sources may complement, while others contradict, each other.

The orthographic transcription of all paired candidate performances added another dimension to the research design, in that selected “episodes” that raters commented on could be closely analysed in relation to the focus of rater comments at different stages of the data collection. Having the fully transcribed paired candidate performances at hand also enabled the identification of exactly which point in the interaction had elicited comments from raters while producing their stimulated verbal recalls (Di Pardo, 1994). The decision to use videotaped rather than audiotaped performances enabled aspects of the performance including body language to be seen by raters, as they would normally have in a live version of the test.

III Rater reactions to an asymmetric interaction

The features that raters regarded positively when rating paired interactions were those associated with Galaczi’s (2004) collaborative pattern of dyadic interaction. These interactions manifested high degrees of equality, as both candidates initiated topics, and mutuality, as both candidates were able to react to their partner’s response by continuing to develop it, resulting in a relatively long topic “life”. It was asymmetric patterns of interaction that proved very difficult for raters, as Galaczi (2004) had predicted. These interactions are characterized by a low level of equality between the participants, and dominant/passive interactional roles as identified by Storch (2001). As Jacoby and Ochs (1995, p. 171) point out, “co-construction does not necessarily entail affiliative or supportive interactions.” This aspect of co-construction is apparent in the speaking test that will be examined in detail. In this test, one candidate dominated the interaction in a way that seemed to the raters to be deliberate and sustained. This caused problems when rating the interactional effectiveness of both candidates, as raters tried to unravel the impact of the dominant candidate on her partner’s opportunity to display his interactional effectiveness.

Galaczi (2004) found that asymmetric dyads were in the minority (10%) of her dataset. Her finding is supported by this study: of the twelve paired speaking tests, only two were characterized as orienting to asymmetry. Both of these interactions were between candidates of similar speaking levels: one between candidates of advanced level, and the other between candidates of intermediate level. However in the context of a high-stakes EAP speaking test, this is a cause for concern, in terms of fairness to candidates.

Paired speaking test: Hu and Ling

This section presents the analysis of three excerpts from a paired speaking test that proved problematic for raters owing to the orientation of candidates to an asymmetric pattern of interaction. By presenting excerpts of candidate discourse in conjunction with the aspects of the interaction raters noted in their stimulated verbal recalls and paired rater discussions, the extent to which raters rewarded and penalised candidates as individuals and as pairs for their role in the co-construction of asymmetric interactional patterns is explored. The asymmetric interaction that will be examined in detail is that between two intermediate level candidates: “Ling”, who is female, and “Hu”, who is male. In the transcription, Ling is “L”, and Hu is “H”. The topic they are discussing is cloning.

Three sequences will illustrate the patterns of interaction that developed as their discussion unfolded. Comments from raters on the particular segment of the discussion as they produced their stimulated verbal recalls will be presented, followed by excerpts from paired rater discussions on this performance.

Sequence 1
H– Hu L = Ling Res = Researcher

H Hao le ma? ((“OK?” in Mandarin))
L ((shakes her head, so Hu continues reading))
Res that’s five minutes
Pause of several seconds

H today we we are talk about this very … have you opinions about the cloning?
L yeah
H ah first one is um banning all the research ah research … what do you think this …this this opinion?
L mm they have both advantages and disadvantages [first]
H [mm]
L I think the advantage is that … er it’s related to the … ethical opinions because cloned animals meet early meet early death and erm so we will find very cruel to kill living things
H so we should ban [the]
L [yeah]
H the human cloning
L ah there are also the disadvantages because cloning ah XX cloning can supply a new way to cure the disease um … especially the cancer cancer is a very terrible disease
H [yes]
L [that] destroys many many peoples for twenty years
H ah yeah this is … ther.. therapeutic cloning
L yes so I don’t agree about this opinion
H yeah

Hu tries to begin the discussion by asking Ling in Mandarin if it is OK to start, but after she shakes her head, he waits for her. After prompting by the researcher to indicate that the maximum five minutes of preparation time is over, Hu once again starts the discussion by asking Ling for her opinion, to which she gives a minimal response “yeah”.

Undeterred by this, Hu continues to manage the interaction by introducing the first proposal – banning all research into cloning – in line 15, and once again asks Ling her opinion. Ling responds with a relatively long turn (lines 19-21) after which Hu paraphrases her point in lines 22 and 24. Ling responds with another long turn (L.25-27), after which Hu clarifies that Ling is actually talking about therapeutic cloning. Ling responds by once again stating her opinion.

At this point in the discussion it appears that Hu is demonstrating quite effective conversation management skills, in that he is able to initiate the discussion, introduce the first proposal, ask for Ling's opinion and clarify a point made her. In contrast, Ling has responded by two relatively long turns and not asked Hu for his opinion, so there are some initial signs of the deeply asymmetric interactional pattern that will emerge as the discussion evolves.

In their stimulated verbal recalls, raters commented on this sequence, and their comments, along with the line number that indicates where they stopped the tape, are given below. Thus L23 means that the rater stopped the tape at the point that corresponds with line 23 in the transcript above. Each rater's comments on the sequence will be presented as one extract.

Extract 2 Rater 1
L23 *also with Ling I'm already thinking her body language isn't very good she isn't maintaining any eye contact she looks very closed off*

L30 *so she's able to give opinions quite clearly*

Extract 3 Rater 2
Prior to the interaction beginning *they seem to be more reticent about starting .. that's initially possibly a problem …..I think they both wish they weren't sitting in this room at the moment … c'mon OK we want someone to take control now*

L15 *alright straight away I'm thinking that Hu's got quite a few problems here … he's having a go but he's not really showing too many skills yet*

L33 *yeah she Ling she's totally controlling the conversation I think Hu is just pretty much along for the ride and that seems to be clear already really*

Extract 4 Rater 4
L 29 *she's not really taking enough notice of him I must say*

While Rater 3 does not comment on any aspects of interactional effectiveness in this sequence, Rater 1 notes that Ling's body language "looks very closed off" and Rater 4 notes that Ling is "not taking enough notice" of her partner. Rater 2 began by commenting on the seeming reluctance of the pair to begin the discussion, which seems a little unfair when it is clear that Hu tries to begin the discussion but is stopped by Ling, and then by the end of the sequence has concluded that Ling is controlling the interaction while Hu is only "along for the ride".

Thus three of the raters have already noted and commented on aspects of the interactional pattern that tend towards the asymmetric at the very beginning of the discussion. One of these aspects is body language. Interestingly, none of the raters noted that Hu had actually displayed a range of conversation management skills, whereas it is evident from the transcribed sequence that he has.

Sequence 2

In Sequence 2 Ling and Hu are in the process of negotiating the relevance of animals to the discussion, as the task sheet had focused on the desirability of pursuing research into human cloning. What transpires in this sequence is crucial to the development of the asymmetric pattern of interaction.

H – Hu L- Ling

L number two is allowing therapeutic cloning but banning re- reproductive cloning … what's your opinion?

H I think er I agree this … er this … this opinion because I think …um … sure ah surely ah therapeutic cloning is er is er is is OK is … ah related to the hu- hu- human illness … but I think … erm the reproductive cloning is the is bad for the people because I think its can … we can talk about it from the three er … er … three aspect first one I think its to ban to ban the the indi- individual … ah … it can make the physical is make your physical is bad … just like the mm just like the first cloning animals the ah the … ah the Dolly … is only …ah …only only alive only alive maybe a month a month

L I think this is a first experience because we lack experience about it … but you know we what you said is advantage about this opinion but I think if we ban reproduc- reproductive cloning

H ah

L a lot of extinct ex- … endangered animals will dying out and er most of them are great fortunate for human

H why do you think?

L from the er … articles we've read

H mm

L about the cloning a lot of species cannot survive for a long time … especially except ah ah ah … a particular species which could survive for thirteen thousand years old

H ahh …

L but most of them are dying out then from ah the generation to generation the species changes is changing

H but I think mm this is talk about the human cloning and not the [animal]

L [this is] not only for the human reproductive cloning is also er refers to the living things

H [yes ah]

L [so esp]ecially for the extinct animals if we have this special advantaged technology I think a lot of animals cannot be ah dying out

H ah …even if these animals be ext- extinct animals was cloned I think this the this cloned animals is not the same as the before … s- … ah s- is not the

L which I agree with

H mm

L is number three

H mm

L ((reads)) "[allowing]

H [right]

L any research related to human cloning which society feel is reas- reasonable and will benefit humanity" … mm the whole society's core are are people

H mm

L it's human society so we should do anything that is benefit for our human ourself then … ah about this research …er also there are many disadvantages and many dis er

H dis

L difficulties

H yeah

L we should er overcome er the first … issues is that the er ethic opinion people think that from the eth- ethic [they will]

H [yes this] is

L [the many]

H [yeah yeah]

L information in the article that cloned animals meet early death very cruel … many people think it's a cruel actions

Ling begins by asking for Hu's opinion, which is the first time that she has done this. Hu responds in a long, disjointed and disfluent turn (lines 35-41), after which Ling follows with a series of relatively long turns (lines 42-55) broken up only by a series of backchannels from Hu.The one time he takes a turn in line 48, it is to ask Ling why she thinks as she does, rather than taking the opportunity to contribute something of his own. It's important to note that Ling allowed Hu to finish his earlier long turn, and did not attempt to interrupt him; however, the disfluency of this rambling turn from Hu may have influenced her later impatience with him.

In line 57 Hu tries to negotiate the topic, as he feels that the focus of the discussion, according to the task sheet, should only be the cloning of human beings, not other animals. Ling, however, maintains that the cloning of any living thing is relevant to the discussion, and proceeds to dominate the rest of this sequence through interruptions. In Line 65 Ling cuts across what Hu is trying to say and changes the topic to a discussion on the third proposal of allowing any cloning research which scientists feel is reasonable. Lines 81-85 see a series of attempts to gain or hold the floor by Ling and Hu, with Ling simply continuing with what she had originally intended to say.

In their stimulated verbal recalls, raters made the following comments on interactional effectiveness in this sequence:

Extract 5 Rater 1
L 39- *OK so already I can hear he's able to provide opinions and give support*

L 41 *OK but I've got a positive response to his body language…he's trying to make eye contact using gesturing*

L 58 *OK Hu tries to go on-track and they respond to each other quite well*

L 63 *OK Hu makes a good attempt at checking meaning with Ling and then clarifies his own idea*

L 71 *and Ling's good at she's very persuasive and very clear about her opinions and getting back on track*

L 79 *OK that was an example of Hu actually encouraging you know Ling in the discussion and even helping her with checking the meaning checking a word*

L 80 *non-verbal skills for Hu are good he's nodding as she's speaking … he's looking at her ..and she has no eye contact.. her body language is not very …doesn't seem to be part of the discussion*

Extract 6 Rater 2
L 71 *I think she's just pretty much ignoring him completely she's not really taking on board what he says .. she's not really making a lot of effort to help him out and try to understand what he's saying but the body language is saying it all really I think .. she's sitting side-on to him she's having the odd facial expression which really says that she's not really interested in understanding what he's saying*

L 84 *yeah there you are she's just speaking over the top of him he's trying to get a word in but he probably doesn't .. well he probably might have something to say but she* *doesn't think it's very valuable .. she's not even really looking at him or even stopping what she's saying*

Extract 7 Rater 3
L 61 *Hu is able to at least question Ling's viewpoint here*

Extract 8 Rater 4
L 51 *now that was quite a good little interaction she gave some you know that was quite a complex sentence and he said why do you think this and she has gone on from there*

L 58 *OK there's another interesting little interaction where they're actually managing quite well .. again this whole thing of we're not talking about animals we're talking about human cloning and he's pulled her up on that and she's managed to assert herself*

L 85 *now he tried to come in there see she's dominating and he tried to come in there but she's just kept talking now I think he needed to assert himself a bit more*

It appears that Raters 1 and 3 view this sequence very differently from Raters 2 and 4. Rater 1's overall impression of Hu's interactional effectiveness is positive, with the features that she notes being Hu's effective body language and his ability to clarify Ling's meaning and notes that they "respond to each other well". Rater 3 makes only one comment during this sequence, and that is to note that Hu is able to "at least question Ling's viewpoint".

Raters 2 and 4 note the increasingly dominant features of Ling's contribution to the discussion, with Rater 2 commenting that Ling "seems to be ignoring" Hu, and noting the way that she was sitting in relation to him, which is interesting as the set up had candidates facing each other, but she has turned "side-on to him", and has "an odd facial expression". The rater infers from Ling's non-verbal communication that she is either "not interested in anything" her partner has to say or "doesn't think it's very valuable". This impression is strengthened by Ling interrupting Hu, with Rater 2 noting "she's speaking over the top of him". Rater 2 appears strongly influenced by Ling's role in the asymmetric pattern of interaction, with overwhelmingly negative comments.

Rater 4 notes the disfluency of Hu's one lengthy turn "now this is painful", but does comment on some "good little interactions" when they negotiate whether other animals are relevant to the discussion. This points to the interaction not being able to be simplistically categorised as "asymmetric" in its entirety, but rather being a series of unfolding mini-interactions, which may display different interactional patterns. By the end of the sequence, however, Rater 4 comments on the manner in which Ling is dominating, but also begins to reflect negatively about Hu.

Sequence 3
H – Hu L- Ling

166 L just now we discuss about three questions
167 H [yes]
168 L [both] the advantages and disadvantages … so you agree about number two or
169 agree about number three … and … ah I think we shouldn't ah pr- … pr- ah ban

170 preproductive cloning because is also benefit for humanity … so I think allow all
171 the research relating to human cloning which society feel is responsible and will
172 benefit human and er … ah … ah new things come from … come to the world
173 will face a lot of difficulties …
174 H yeah
175 L if you want to go ahead go too deep and ah which is benefit for the human we
176 should encourage them
177 H mm
178 L and give them more incentive … … and so … the more important thing is that the
179 whole thing the core of the society is people … ar people it's person so we
180 should try our best …
181 H we should try and encourage them
182 L so I think think we also to talk about our own opinions … ah this is issue … ah
183 with the er… tech … with the technology growing it has the way to ah … deal
184 with these issues … ….
185 H yes I think erm … I think ah … human cloning erm … does if ah … if ah … if
186 according to what you said I think it's right
187 L … … but if I think if you don't ban to reproductive cloning there's a new field
188 H mm
189 L for society to explore … I think it's more benefit for human society … that's all

Ling begins to summarize the discussion in Line 166, and actually does try to elicit Hu's opinion in Line 169, but after a slight pause, continues with her turn, which could indicate a number of possibilities: she is not expecting Hu to respond in a meaningful way, is unprepared to wait for his response, or is simply uninterested in anything that he might have to say by this stage in the discussion.

This is followed by a series of relatively long turns by Ling, with Hu giving minimal backchannels such as "yeah" and "mm", and echoing Ling's point in Line 181. When Hu does try to contribute his opinion in Lines 185-186 the disfluency is clear.

Ling concludes with an "I" statement in Line 189, indicating that she is not representing a shared view, nor has a meaningful consensus of opinion been established. It is Ling who ends the discussion decisively with "that's all", indicating once again the non-negotiable control that she has taken over this discussion.

From the stimulated verbal recalls, raters responded to this exchange in the following way:

Extract 9 Rater 1
L 187 *that was good Hu attempts to agree with .. makes consensus*

L 189 *OK so she sums up well*

Extract 10 Rater 2
L 164 *yeah he's not responding at all to her questions now .. he didn't understand that even after she explained it again .. so he's slipping back and back really*

L 187 *there "everything you said is right" very good ((sarcastic tone))*

Immediately following the end of the performance: *poor Hu*

Extract 11 Rater 4
L166 *now I think I've overestimated him because he's just not producing the goods … that's pretty pathetic ..his repetitions you know he's just not up to it really*

L 169 *and she's just plowing on like a bus*

L 187 *now he's basically given up I mean he's just been overwhelmed by her stream of words and whereas in the middle he actually had periods where he actually could express himself I think he's with a he's mismatched or he's in one of those nightmare situations where she's taking almost no notice of him and I think it's destroyed his confidence and he's hesitant anyway and this hasn't helped*

L 189 *and the interesting thing is that she decides when to end the whole thing as well she's just she's basically totally in charge of this interview and he's been left way behind*

Once again it is striking that Rater 1 continues to comment positively about Hu, and merely notes that Ling "summed up well". At this stage she seems not to have noticed the asymmetric nature of the interaction. Rater 3 makes no comments on this section, so we can only assume that she does not perceive the domination either. Yet both Raters 1 and 3 comment on this during the paried rater discussions, which I will come to shortly.

Just prior to this sequence beginning, Rater 2 noted that Hu has "stopped responding at all to her questions" and is "slipping back". The only other comment he makes as the sequence unfolds is a rather sarcastic echoing of Hu's "everything you say is right", and an empathetic "poor Hu" at the end of the performance.

It is Rater 4 who seems most attuned to the asymmetric nature of the interaction, and the impact it is having on Hu. Although she comments rather unsympathetically at the beginning of the sequence "now I think I've overestimated him…he's not up to it really", she later reflects "I think he's … mismatched or he's in one of those nightmare situations where she's taking almost no notice of him and I think it's destroyed his confidence ". She refers to Ling as "plowing on like a bus", with Hu "overwhelmed by her stream of words" and these graphic images do seem to reflect the actual discourse of the candidates in this sequence. Rater 4 even refers to the discussion as an "interview" in her final comment, which shows the extent to which the asymmetric pattern was established.

Rater 4 gives Hu a Band 2 for Effectiveness in this pairing, but when he is paired with a female candidate of a higher oral proficiency level and does a parallel discussion task, he not only contributes more (53% as opposed to 38% when paired with Ling) to the interaction that is perceived as collaborative by all raters, he is given a Band 4 for Effectiveness by Rater 4. This is potentially an issue of concern in any test, but particularly so for a high-stakes test.

After they had produced the stimulated verbal recalls, raters came together in pairs to discuss their ratings. The following are excerpts from the discussion between Rater 1 (R1) and Rater 2 (R2). Extract 12 begins with Raters 1 and 2 realizing that they had very different impressions of the two candidates' interactional effectiveness, which was predictable when we consider the comments they made in the stimulated verbal recalls.

Extract 12 Raters 1 and 2

R2 *well she was kind of closed from the beginning*R1 *yeah so I I was feeling a bit for him there with*

R2 *but he had no assertiveness whatsoever*

R1 *no he was trying to engage with her and trying to*

R2 *but when she gave him the talking stick he just couldn't express himself at all*

R1 *mm I didn't think he was so bad*

[Later in the discussion]

R1 *I just felt…she's saying a lot of information but not* <u>*to*</u> *him or asking him what he thinks…so I mean in a way you can't judge someone on anything other than what we see but it seemed to me that he was poorly affected by her lack of interaction*

R2 *I thought if anything that she was the one affected by him ..and so…..really it was a question of him holding her back ..because he gave her nothing to work with..and she couldn't really respond coherently because he wasn't really making many points…she was the one making the points…I think with a better partner she might have seemed better so I would go up to a 3.5 for her*

[Later in the discussion]

R1 *I think for example if she were at university*

R2 *you think she'd be alright*

R1 *she was interpreting texts ..she was using them as supporting information… even though she had these intonation problems I think it didn't impede communication ..she was able to express her ideas she was quite persuasive …as you said she was able she was leading and he was agreeing with everything she said and*

R2 *so 3.5 for her*

R1 *can we go to 3.75*

R2 *ah:: alright because then it comes down to that she's still not a pass but I don't know the extent to which he held her back*

From the discussion between Raters 1 and 2, the difficulty that the raters feel in coming to a rating decision that is fair to both candidates is clear. Rater 1's comments reflect the dilemma that she feels in that her training tells her that she can only base her decision on what Hu has produced, not what she feels he might be capable of with a different partner. Her understanding of the interaction is that Hu's performance was negatively affected by Ling's domination and lack of genuine interaction with him.

Rater 2, however, reaches the opposite conclusion and decides that it is Ling who has been adversely affected, and explicitly states that he would be willing to compensate for this in his score for effectiveness. He also speculates on what Ling might be capable of doing with another partner. Rater 1 then speculates as to whether Ling could cope with the speaking demands of university study, and this seems to influence both raters, who finally decide on giving her a 3.75.

Considering Rater 4's comments during the stimulated verbal recall, it was to be expected that she would pursue the impact of the Ling's domination in the paired rater discussion she had with Rater 3. The following excerpt is taken from the discussion between Rater 3 and Rater 4 as they focus on how to deal with this interaction in terms of their ratings.

Extract 13 Raters 3 and 4

R4 *at one point he sort of lost his confidence really it was like*
R3 *which was no wonder she didn't*
R4 *that's what I mean*
R3 *she didn't give him much chance to*
R4 *yes if he'd been you know drawn out a bit more I have a feeling he could have managed a whole lot more*
R3 *he didn't have much chance to say anything*
R4 *no so I think she should be penalized for effectiveness and for him as well probably what do you reckon I mean he did say a few things to her "do you think"*
R3 *yeah*
R4 *he tried he tried*
[Later in the discussion]
R4 *you've got to keep your emotional judgement out like if you feel slightly I mean I was so not actually hostile but I was extremely aware of her you know unwillingness to include him*
[Later in the discussion]
R4 *look if she'd have had someone who was equal to her it might have really brought her out a bit more too and you know forced her to think about things a little bit more you know and held her back a little bit I don't know but just the lack of balance between the two of them*
R3 *it disadvantaged him more than anything*

In the extract above, Raters 3 and 4 discuss who was disadvantaged by the asymmetric pattern of interaction, who should be penalized for it, and how this should be reflected in the scores that they award. They note the extent to which Ling seemed to deliberately intend to dominate, and Rater 4 reflects on the dissonance that she feels. This illustrates the complexities inherent in rating a paired candidate oral performance, particularly when things are not going successfully. While Raters 3 and 4 point out that if Ling had been with a "stronger" partner, she may not have been allowed to dominate, and indeed may have been extended if she were "forced to think about things a little more", their sympathies lie with Hu, with the final comments from them being "the lack of balance between the two of them…it disadvantaged him more than anything". Thus the two pairs of raters interpret the cause and the impact of the asymmetric interaction differently, although this is not fully reflected in the scores they award.

To make inferences purely on the basis of scores given for interactional effectiveness would be misleading, as data from rater notes, stimulated verbal recalls and paired rater discussions indicate similar scores may not be the result of the same rating decisions when explored more deeply. The difficulty that raters had in reaching a decision on interactional effectiveness, particularly when interactional patterns were asymmetric, is also not fully reflected if we consider only quantitative data. In addition, the impact of the pairing of candidates, while mentioned in the stimulated verbal recalls, most explicitly surfaces in the paired rater discussions.

IV Conclusions

It is clear that paired speaking tests have the potential to elicit features of interactional competence, including a range of conversation management skills, that are generally not featured in traditional language testing interviews. The features salient to raters in this study are best captured through tasks that involve candidate-to-candidate interaction. If we value these skills, they should be incorporated into task design and rating scales. However, it is the operationalisation of interactional competence that is a concern.

Thus on a practical level, there are implications for the development of rating scales that more accurately and thoroughly encompass the complexities of interactional competence in a paired speaking test, and the training of raters to deal with asymmetric interactions. That raters noted and incorporated body language, assertiveness through communication, the ability to manage the discussion and work together cooperatively into their frames of reference for interactional effectiveness attests to the complexity of the decisions they are required to make, and the lack of meaningful descriptors in guiding them to this decision. Raters particularly need direction in dealing with asymmetric dyads, especially where one partner deliberately dominates the other. If individual scores are to be given, there is a clear need for a policy of either compensating, or "no benefit" to candidates who are perceived to have been disadvantaged by the interactional style of their interlocutor. The rationale for this decision would reflect the construct of interactional competence that the test developer is operationalising.

When raters in this study were faced with trying to unravel the extent to which one candidate's interactional style had impacted on his/her partner, they would consider two factors: the potential of the candidate if s/he had been with a different partner, and the Target Language Use situation- a university seminar or tutorial. The dissonance that raters feel when faced with this situation is evident when Rater 1 says: *"so I think in a way I mean you can't judge someone based on any other thing that what we see but it just seemed to me that he was poorly affected by her lack of interaction".*

In ethical terms, the fairness of exposing a candidate to a test format that may disadvantage him/her because of the interactional style and comprehensibility of their partner is an issue of concern, as is the role that a candidate may have to take on when paired with a candidate who is weaker in interactional skills, which may not always have a direct correlation to the candidate's overall spoken proficiency. The consequences for a candidate involved in an asymmetric interaction are very real, not simply a matter of rater perception. In the example explored in this article, a candidate is given a score of 2 by a rater when involved in an asymmetric interaction with a partner of a similar level. When this candidate interacted with a candidate of a higher level, and the interaction was seen by the raters as being collaborative, the same rater awarded this candidate a score of 4. Thus it is important to acknowledge that interactional patterns established by interlocutors in one pairing were not necessarily carried through to another pairing. This points to the local and situated nature of each paired speaking test.

In addition to the level of spoken proficiency, other factors related to personality and culture might determine, or at least contribute to, the pattern(s) of interaction that emerge in a paired candidate speaking test and thus influence raters' perception of a candidate's interactional effectiveness. That both body language and demonstrating assertiveness through communication were salient to raters in determining the effectiveness of a candidate's interaction may be of concern, in that these characteristics could be seen as aspects of culture, and L1 usage. What constitutes assertiveness in communication to the raters might be interpreted as aggressive and inappropriate communication in some cultures. This point has also been made with respect to turn-taking by Fulcher (2003, p. 35): “The turn-taking routines and conventions that apply to Anglo-American societies do not apply to all societies”.

It is clear that separability is an issue of concern in paired speaking tests, particularly in terms of the co-construction of interactional patterns which may be perceived as being disadvantageous to candidates. The pattern of interaction characterised as “dominant/passive” by Storch (2001) and “asymmetric” by Galaczi (2004), caused raters difficulty in trying to unravel the impact of one candidate upon another, in order to award separate- and fair- scores for interactional effectiveness. Raters had an inherently social view of interaction, and thus perceived the interactional patterns that dyads oriented to as co-constructed. Some features of interactional effectiveness are clearly perceived by the raters to be mutual achievements of the two candidates. These features include mutual comprehensibility, the ability to respond effectively and the authenticity and quality of the interaction. If these aspects are regarded as mutual achievements, it seems counter-intuitive to ask raters to award separate marks to candidates for interaction. Are these features then to be ignored, or factored out of ratings- and the task design? One possibility could be to award a shared score to paired candidates for interactional effectiveness, while still awarding a separate score for Accuracy, Fluency and Range. Swain (in Fox, 2005, p. 242) also suggests this: “as far as evaluating the individual in an interactive test, I’m not sure whether it makes sense to do it”.

If a decision were made to award shared scores, the next challenge facing test developers would be deciding exactly what the descriptors would need to encompass. In order to recognize the co-construction of the interactional patterns, descriptors for shared scores could draw on the features of the collaborative, blended, parallel and asymmetric interactions, which have been identified by Galaczi (2004). For example, a descriptor for a shared high score for interactional effectiveness could include: “the candidates oriented toward a collaborative pattern of interaction, characterized by high levels of equality and mutuality”; whereas a descriptor for a shared low score for interactional effectiveness could include: “the candidates oriented toward an asymmetric pattern of interaction, characterized by low levels of equality, and a monopoly on topic initiation and development by one speaker”.

Yet while it is clear that the raters valued – and rewarded- candidates who oriented to relatively symmetrical, collaborative patterns of interaction, the question remains as to whether collaborative interactional patterns should be uncritically positioned as the “gold standard” of communication: it could be argued that many of the interactions that take place in a university context are by their very nature asymmetrical; for example, a student requesting an extension on an assignment from a professor. In this respect, the limitation of using only one task is apparent, as we cannot generalize to other communicative contexts and purposes, where a student might be required to achieve their communicative goal when faced with a more powerful interlocutor. Such a situation of inherently asymmetrical interaction might approximate more closely to a traditional language testing interview- or a role play with roles that clearly set up an unequal relationship between interlocutors. Further research including different tasks in paired candidate speaking tests, including role-plays, would be valuable to pursue. In addition, the question of whether it is possible to establish a truly collaborative interaction in a speaking test remains a challenge to the field. In the context of a high-stakes test, it seems somewhat disingenuous to expect candidates to naturally orient to a collaborative interaction and display a wide range of interactional competencies, as He and Dai (2006) found in their study of group speaking tests.

While this study was able to examine the rater's perspective on interactional effectiveness and integrated several extracts from the candidate's discourse, Galaczi's (2004) study has shown the value of a full CA of candidate discourse. Another potentially enriching perspective would be that of the candidates themselves, which could be gained through the use of stimulated verbal recall. This would provide an insight into how the candidates framed the task, and whether in the asymmetric interactions the candidates may have been involved in the "same task, different activity" in that each candidate may have understood the communicative purpose of the test in quite different ways. Further research may also involve the impact of the inclusion of a trained interlocutor who can intervene in paired speaking tests, as is the practice in University of Cambridge ESOL tests. While this may mediate against candidates being disadvantaged by being with a partner who orients toward asymmetric interactional patterns, the impact of the interlocutor on the interaction and the way that this is taken into account in the ratings would be valuable to ascertain.

In conclusion, the research findings suggest that the issue of the separability of individual candidate's contribution to an asymmetric interaction in a paired speaking test is a concern, in terms of both construct validity and fairness. Whilst I would suggest that shared scores for interactional effectiveness in a discussion test task would, at this stage of our knowledge of the construct, seem to be one way of acknowledging the essential co-construction of the interaction, it is difficult to see how this could be implemented, particularly in the case of high-stakes tests, where there is a practical imperative for the desire to have neatly separable "performances", as McNamara and Roever (2006, p. 51) point out: "institutional needs are in line with the psychometric orientation to individual cognitive ability: what is required is not a faithful account of the interaction but a score about individual candidates that can then be fed into the institutional decision-making procedures". This dilemma, as Chalhoub-Deville (2003, p. 373) identified, is inherent in trying to reconcile "the notion that language ability is local" with the "need for assessments to yield scores that generalize across contextual boundaries."

References

Brown, A. (2000). An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (Ed.), *IELTS Research Reports, 3*, 49-84.

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

Brown, A. (2004). *Interviewer variability in oral proficiency interviews.* PhD thesis. University of Melbourne.

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

Butler, F.A., Eignor, D., Jones, S.,McNamara, T. and Suomi, B.K. (2000). *TOEFL 2000 Speaking Framework: A Working Paper*. TOEFL Monograph No. MS-20.

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. Eotvos Lorand University, Budapest.

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.* Thousand Oaks, CA: Sage.

Egyud, G. and 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. (2005). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly, 1, 4,* 235-251.

Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Pearson.

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

He, A.W. and Young, R. (1998). Language Proficiency Interviews: a discourse approach. In Young, R. and He, A.W. (Eds.) *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamins.

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

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. and Ochs, E. (1995). Co-construction: an introduction. *Research on Language and Social Interaction, 28, 3, 171-183.*

Kramsch, C. (1986). From Language Proficiency to Interactional Competence. The Modern Language Journal, 70, iv, 366-372.

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.

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. and 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. and Roever, C. (2006). *Language Testing: The Social Dimension.* Malden, MA: Blackwell Publishing.

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

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

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.

Weir, C.J. (2005). *Language Testing and Validation*. New York: Palgrave Macmillan.