Raters' orientation in a paired speaking test

Sayamon Singto[1]
Department of English, Thammasat University, Bangkok, Thailand

This qualitative research study examines the ways in which raters assessed the performance of students in a paired speaking task. The research objectives were to study 1) how the raters assigned scores to the students' performance, and 2) the features of interaction that the raters attended to when assessing students' proficiency in a paired speaking test. Seventeen raters were asked to watch three video clips of three different pairs of students performing a paired task. The raters were subsequently asked to assign scores and provide comments on the features of interaction that they attended to while watching the clips. Their comments were recorded and later transcribed for the analysis. Results indicate all of the raters in the study assigned scores to each individual student rather than as a pair. With the exception of one rater who used analytic rating scales, the remaining sixteen employed holistic rating scales. The analysis of their comments suggest five main categories of interaction that the raters attended to when assessing the students' performance,

[1] Sayamon Singto is a lecturer in the English Department at Thammasat University. She holds a PhD in Language Education from the University of Georgia, United States. Her interests include classroom interaction, Conversation Analysis (CA), and assessment of speaking proficiency. She may be reached at singtosayamon@gmail.com.

including *task completion, accuracy, interactional management, naturalness, and fluency.*

Keywords: paired speaking test; rating criteria for speaking test; raters' orientation

งานวิจัยเชิงคุณภาพชิ้นนี้มุ่งศึกษาแนวทางที่ผู้ประเมินใช้ในการประเมินผลการสอบทักษะการพูดแบบเป็นคู่   โดยมีจุดประสงค์เพื่อศึกษา 1) วิธีการให้คะแนนของผู้ประเมิน และ 2) ลักษณะต่างๆของการสนทนาที่ผู้ประเมินใช้ในการประเมินผลการสอบทักษะการพูดแบบเป็นคู่ โดยใช้ข้อ มูลจากผู้ประเมินทั้งสิ้นจำนวน17 คน ที่ได้ดูวิดีทัศน์การสอบทักษะการพูด ของนักศึกษา 3 คู่ หลังจากดูวีดีทัศน์ของนักศึกษาแต่ละคู่เสร็จสิ้น ผู้ประเมินจะทำการให้คะแนน และพูดอธิบาย เกี่ยวกับลักษณะต่างๆ ของการสนทนาที่ผู้ประเมินใช้ในการให้คะแนน รวมทั้งเหตุผลและ เกณฑ์ในการให้คะแนนกับนักศึกษา ในระหว่างการบรรยาย มีการบันทึกเสียงของผู้ประเมิน ซึ่งได้นำมาถอดเสียงถ้อยคำ เพื่อนำมาใช้ในการวิเคราะห์ผลผลการวิจัยพบว่าผู้ประเมิน ทุกคนประเมินผลการสอบของนักศึกษาโดยแยกเป็นรายบุคคล ไม่ได้ประเมินผลเป็นคู่ และผู้ประเมินส่วนใหญ่คือ 16 คน ใช้วิธีการให้คะแนนโดยใช้มาตรวัดประเมินค่าแบบภาพรวม ส่วนอีก 1 คนที่เหลือนั้นใช้มาตรวัดประเมินค่าแบบแยกประเด็นส่วนลักษณะต่างๆ ของการสนทนา ซึ่งผู้ประเมินใช้เป็นเกณฑ์ในการให้คะแนนสามารถแบ่งเป็นกลุ่มหลักๆได้ 5 กลุ่ม อันได้แก่ *task completion, accuracy, interactional management, naturalness,* และ *fluency*

คำสำคัญ: การสอบทักษะการพูดแบบคู่ เกณฑ์การประเมินผลการสอบการพูด แนวทางการประเมินผลสอบของผู้ประเมิน

Raters' orientation in a paired speaking test

Sayamon Singto
Department of English, Thammasat University, Bangkok, Thailand

Introduction

Interaction-based tasks such as pair or group work have increasingly been used in classroom and assessment contexts in response to the move toward a more communicative approach in language teaching (Iwashita, 1997; Taylor, 2001; Taylor & Wigglesworth, 2009). When incorporated in low stakes classroom assessment context, it is hoped that these tasks provide a positive washback effect for the classroom, giving students additional incentive to collaborate with one another when participating in classroom activities. A pair speaking test format has also been included in high stakes assessment such as Cambridge ESOL General English Main Suite to counter criticisms directed toward an examiner-test taker interview format. The growing popularity of pair assessment tasks in both contexts has stimulated research examining how this form of assessment might affect test takers and assessors. Second language acquisition research has provided evidence supporting the use of this test format. When compared to an examiner-test taker interview format, peer-to-peer interaction provides opportunities for more substantive and genuine interaction, enabling interlocutors to produce a greater variety of language functions (Brooks, 2009; Gan, 2010; Kormos, 1999; Macqueen & Harding, 2009; Saville & Hargreaves, 1999; Swain 2001), and the paired format is viewed positively by test takers as it is considered less stressful (Együd& Glover, 2001; Ikeda, 1998). However, Foot (1999) questioned whether test takers' preference alone can justify removing the examiner-test taker interview format when certain aspects of a pair format can potentially be unfair to some candidates. The task of judging test takers' performance in a pair test is a complicated one, even when raters are given a set of criteria on which their decisions about test takers' abilities will be based. Of course, without specific criteria, the task seems even more daunting, and the fairness and reliability of scores are questioned.

This is precisely the challenge faced by raters participating in this study. When these raters, who are also course instructors, are asked to teach the class, they are provided with a course outline that does not specify assessment criteria. In this circumstance, they are compelled to draw on their interpretation of the course objectives, their concept of what oral proficiency entails and perhaps even their intuition to construct their own assessment criteria. The present study is motivated by the need to identify features that raters take into account when assessing a paired task based on the comments they provide and the scores they give to test takers. In short, this study explores how raters assess students' performance in a paired assessment task when no specific criteria are provided.

Review of literature

Test takers

The increasing popularity of paired assessment tasks has prompted researchers to examine various aspects of the tasks, shedding light on both the positive outcomes and problematic issues resulting from the use of this oral assessment format. An area that has received considerable scrutiny concerns the interlocutor effects, as in a context of paired oral assessment in which test takers' performance is co-constructed—one test taker's engagement in the task is, to a certain extent, affected by the other, in positive and negative ways. Findings from research in this area indicate that proficiency level, degree of acquaintanceship, as well as personality appear to influence the scores and the amount and complexity of talk produced.

It has been hypothesized that the pairing of test takers with differing degree of oral language proficiency may benefit one test taker but disadvantage the other, prompting concerns about the fairness of this assessment format. For the most part, research findings suggest that test takers' different level of oral proficiency does affect the nature of talk, but it does not always impact their scores. Test takers with lower level of proficiency tended to talk more when they were paired with an interlocutor of higher proficiency (Davis, 2009; Norton, 2005). However, in Davis' study the amount of talk did not correlate with the scores given. I washita (1996) also reported similar findings that test takers produced more talk when they were paired with an interlocutor with a higher level of proficiency; however, contradicting results were found with regard to scores. Specifically, test takers with lower levels of proficiency appeared to benefit from being paired with more proficient partners as they were awarded slightly higher scores.

In addition to level of proficiency, other factors that have been associated with test takers' performance include familiarity and personality. The notion of acquaintanceship has been considered to have an impact on the nature of test taker interaction, as well as the outcomes or scores given. O'Sullivan (2002) found that female test takers achieved higher scores when they were paired with a friend than with a stranger. Yet, when examining the accuracy and complexity of language produced, no significant difference was found. Aside from test takers' familiarity with one another, individual personality traits are assumed to impact their performance and scores. The effect of test takers' personality was explored in Ockey's (2009) study in which assertive test takers were grouped either with other assertive test-takers or with non-assertive ones. The researcher found that assertive test takers only scored higher than expected when grouped with all non-assertive test takers, but not when all other group members were also assertive. In addition to assertiveness, extraversion has been found to have some effects on discourse in a group oral test (Nakatsuhara, 2011).

Another line of investigation focusing on test takers seeks to explore possible relationship between assigned scores and discourse patterns. Lazaraton and Davis (2008) employed Conversation Analysis (CA) to analyze high-scoring test taker interactions in Cambridge ESOL Speaking Tests, with the aim to demonstrate different ways in which test takers positioned themselves as proficient language users in their discourse. The detailed CA transcripts clearly illustrated features of discourse that account for high scores both in cases in which test takers were equally proficient and interactive, and those in which there was a mismatch in proficiency and interactivity. The former case, which occurred when test takers were on par with each other in terms of proficiency and interactivity, tended to produce collaborative talk that merited high scores. In the latter case where proficiency mismatching occurred, higher scoring ones in the pair managed to use discourse moves to cast themselves as a supportive partner or to display assertiveness, earning themselves relatively high overall scores. Similar findings were reported in Galaczi's (2008) study, whose main focus was to identify overall patterns of interaction in peer-peer interaction in Cambridge ESOL Speaking Test. The researcher found that dyads whose interactions were characterized as "Collaborative" earned highest scores. Two other patterns of interaction identified in this study were labeled "Parallel" and "Asymmetric". Characterization of these patterns of interaction was based on three underlying features: *mutuality*, which referred to coherence of one turn of talk and the next; *equality*, which involved distribution of turns of talk between the test takers; and *dominance*, which encompassed quantity of talk, interruption and questions.

Raters

While the focus thus far has been on test takers, another significant component in the assessment process is raters. Research studies on raters generally examine raters' attention to various features of interaction, their scoring decision, and the extent to which they viewed the interaction as co-constructed. These issues are pertinent to the validity and fairness of scores. Orr (2002) compared raters' scores and their comments of the same performance and found that raters held different interpretations of the criteria, resulting in contradictory perceptions of the same performance. The first outcome was that the same performance was awarded the same score by raters, but their comments on the performance varied greatly. For instance, when two raters rated a performance as 3, representing a "simple pass", on the construct of grammar and vocabulary, one rater commented that the test taker "had a noticeable problem with pronouns" while another rater's view of the same test taker was clearly more positive, noting that "There was good use of pronoun" (p. 146). The alternative was that the nature of raters' comments for the same performance were quite similar, yet different scores were given. The researcher also found that raters paid attention to other non-criterion features such as how the test takers presented themselves and how they fared compared to others. Similar findings were presented in May's (2006) study, in which raters mostly adhered to features listed in the criteria, but also incorporated features that were not explicitly mentioned in the criteria and those that were non-criterion aspects of test-takers' performance, which were categorized as "Rater Reflection" and "Task Realization". In keeping with these findings, Pollitt and Murray (1996) found that some of the raters in their study seemed to form a holistic image of the speaker based on a few first impressions that served as a primary indicator of level, and some of these raters' comments were not based on observable performance data.

Yet another problematic issue arises when raters make judgment about test takers' interactional competence. Raters are influenced by their views of what constitutes a successful interaction. In any performance-based assessments that involve interaction between individuals, the discourse is jointly constructed. In other words, the completion of the task requires contribution from each interlocutor. As such, McNamara (1997) questioned whether "communicative competence can be viewed as residing in the individual" (p.457), and whether any one interlocutor could be held accountable for problems or communication difficulties. May (2009) explored how raters assigned scores for interaction effectiveness to individual test takers in the interactions that were labeled asymmetrical. Like raters in Galaczi's (1998) study discussed above, May found that collaborative patterns of interaction were viewed positively by raters, while asymmetrical patterns of interaction presented difficulty to raters in giving scores that they felt were fair to each test taker. Raters' comments clearly showed that they took into consideration the impact of dominant test takers on their partners and vice versa. As a result they either compensated or penalized test takers for their roles in the interaction. Based on her findings, May suggested that test takers interaction be viewed as mutual achievement and that a shared score for interaction effectiveness be given to both test takers instead of rating them separately. In her later work, May (2011) identified interactional features that raters associated with higher and lower scores for interactional competence. An analysis of raters' comments also indicated that although raters were trained to rate candidates separately, there were certain interactional features that raters were likely to consider as mutual achievements. Given the focus on the construct of interaction, Ducasse and Brown (2009) identified three main components in test takers' interaction that raters associated with successful interaction including: non-verbal interpersonal communication, interactive listening, and interactional management.

Scoring speaking test

Scoring a speaking test involves three main issues: assessment criteria, rating scales and types of scoring. Assessment criteria reflect test developers' beliefs and assumptions about the nature of language, and they are important tools to distinguish good performances from weak ones. For instance, the overall criterion for Test of Spoken English (TSE) is communicative effectiveness and the four analytical criteria are functional, sociolinguistics, discourse, and linguistic competence (Douglas and Smith, 1997). Similarly, the speaking part of Cambridge ESOL General English Main Suite, which includes a collaborative task between two test takers, awards one mark for overall effectiveness and one each for three, four or five analytical criteria depending on the level of the exam. For the most basic level in the suite of exams, Key English Test (KET) performance is assessed based on three criteria: *grammar and vocabulary*, *pronunciation*, and *interactive communication* (UCLES, 2008a). In the Certificate of Proficiency in English (CPE), the highest level of the exam, the criteria are *grammatical resource*, *lexical resource*, *discourse management*, *pronunciation,* and *interactive communication* (UCLES, 2008b).

Once assessment criteria have been defined, two other issues to consider are types of scoring and rating scales. Two main approaches in assigning scores include holistic and analytic scoring. Holistic scoring results in a single score that represents an overall impression of test takers' performance. Rating scales for holistic scoring, therefore, contain short descriptions of different levels of language ability encompassing all the criteria. On the contrary, when analytic scoring is used, test takers receive a set of scores, one for each of the criteria used (e.g., accuracy, fluency, pronunciation, etc.). Analytic rating scales usually contain 3-5 criteria, each with its descriptors at different level of the scales (Luoma, 2004). In addition to these two types of rating scales, there exists a diagnostic rating checklist, which, as its name suggests, contains a list of features of successful performance of a particular task. Raters use these lists as they observe test takers' performance, making note of features that are present or absent in test takers' performance. These checklists can be used on their own or in combination with holistic or analytic scales.

To conclude, previous work on various aspects relating to assessment of interactional tasks has provided insights into different approaches that raters use to judge students' performance and factors that can possibly influence raters' decisions. Although the present study does not focus on the test takers, studies focusing on test takers engender an understanding of the interplay of test takers' characteristics, interaction patterns, and resulting scores.

Research questions

This study seeks to explore the approaches that the raters use in assigning scores and the raters' orientations to various features of students' performance in a paired speaking task. Therefore, two main research questions govern this study:

1. What approaches did the raters use in assigning scores?
2. What features of performance did the raters attend to when assessing students' proficiency in a paired speaking test?

Methodology

The course

The context of this study is a beginner level English conversation course in a Thai university. The course is compulsory for all English majors and minors. Students from several other disciplines take this course as one of their English language course requirements. A smaller number of students take the course as an elective. The course is taught either by native speakers of English or Thai instructors. Two main objectives of the course are 1) to help students communicate in English with the basic situations of everyday life using language functions and other features of spoken English with an emphasis on those situations the students may face in Thailand, and 2) to help students speak English more confidently, fluently, accurately and appropriately. As stated in the course outline, the communicative approach is adopted, and English is the medium of instruction, although this is not stated directly in the course outline. One of the typical methods of assessment for the course is a role-play between a pair of students based on a given situation that reflects language functions covered in the course. The course outline does not indicate particular criteria that instructors have to use in assessing students' proficiency. This is also the reason participating raters were not given any rating criteria when they were asked to assess students' performance.

The raters

Virtually all instructors who had taught the course prior to the time of data collection were asked to participate in the study. The 17 instructors who agreed to participate in the study were full-time instructors at the time of data collection. Nine of them were non-native speakers of English, and the remaining were native English speakers.

The paired speaking task

The speaking task that was the focus of this study is a role-play based on the contents from the first half of the course. This task format is commonly used in class to practice various language functions and for assessment. The researcher was granted permission to record role-plays in an actual test situation in one of the classes from a semester prior to the time of data collection. Video recordings from three pairs of students were chosen for use in the present study as they represented typical pairing types that occurred in the context of this course. In most classes, instructors allowed students to choose their test partners, so pairs tended to vary with respect to level of proficiency and gender.

In the test situation featured in the three video recordings chosen for the present study, each pair of students was given a card describing a situation in which the conversation was to take place, the language functions required within the conversation, and a few "precautions" from the instructors (See Figure 1). They were given a few minutes to prepare for the task together. The italics show parts of the tasks where variation of topics occurs. For example, instead of making a plan to go to a concert as illustrated, students may be asked to make a plan to go to a movie. The situations were randomly chosen for each pair of students; therefore, the students did not know exactly what they would be asked to talk about beforehand, although they did have a general idea about the task based on their experiences in class.

Situation: At a party you meet each other for the first time. It seems that you get along very well so you make a plan to go to a concert together one evening next week.

Language functions to be covered:

- Greeting & Introducing yourself to each other
- Asking and answering questions about personal information (showing interest in different ways)
- Talking about likes, dislikes and preferences about music.
- Agree/Disagree with likes and dislikes
- Making suggestions about a possible concert to go to
- Accept/Reject suggestions
- Making an appointment (what time and where to meet)
- Leave-taking

Cautions:

- Your conversation should last about 2 minutes.
- Do not spend too much time on one language function.
- Do not look at your notes or textbook.

Figure 1. Situation card.

Data collection

This study employed retrospective verbal protocols to capture the raters' attention to various features of the paired task that were taken into account when assessing students. The raters were asked to assess students' performances using their own scoring approach and criteria. They viewed video clips of three different pairs of students performing a speaking task described above, each of which was shown only once in its entirety without pausing. Subsequently, the raters were asked to give an assessment of students' performances and comment on the features that they attended to while watching the video clips. Their comments were recorded and later transcribed. For the non-native raters their transcripts were translated into English for the purpose of the analysis. As the main objective of this study was to explore how the raters assessed students' performance in the context of this course, the raters were neither instructed how to assess students nor given any criteria. Instead, they were asked to proceed as they would when assessing students in their own class. The 51 verbal reports from 17 raters commenting on 3 pairs of students constituted the primary data for this study.

Data analysis

The 51 verbal reports were transcribed, and each transcript was examined to identify the scores that each rater gave to test takers. Following the verbal protocol analysis procedure suggested in Green (1998), the transcripts were divided into units for analysis consisting of a single or several utterances with a focus on a single event or idea.

As such, repetitions and elaborations were not considered as new units. These units were then coded into categories representing features that the raters attended to in their assessments of students' performances. Multiple categories that emerged in the coding process were grouped by theme to form main categories of comments[2].

Findings and discussion

1. What approaches did the raters use in assigning scores?

The table below shows the scores that each student received. With the exception of rater 7, all other raters employed holistic scoring and awarded a single letter grade or score for the performance (see Table 1). Rater 7 used analytical scoring based on five criteria: "communicative competence", "language use", "pronunciation", "fluency" and "naturalness", each receiving 10 points (See Table 2). Raters 10 and 11 awarded numerical scores that were based on the total scores of 10 and 20 respectively. For the purpose of comparison, these numerical scores were converted to their equivalent letter grades. The conversion was based on the general grading criteria of the course (i.e., A = 90%, B+ = 85%, B = 80%, C+ = 75%, C = 70%, D+ = 65%, D = 60%, and below 60% = F). It is important to note that some of the raters, regardless of the types of scoring approaches they used, occasionally

[2] The coding process might alternatively be described within the framework of grounded theory: opening coding, axial coding and selective coding (see Glaser and Strauss, 1967).

assigned a grade range instead of a single letter grade or numerical score (e.g., rater 2, using holistic scoring, put students C grade into a C to C+ range, and rater 7, using analytical scoring, gave the same student a range of 6-6.5 for the "language use" and "fluency" criteria). In this respect, the findings concur with those from previous research suggesting that fuzziness in assessing speaking performance can occur regardless of the types of scoring and rating scales the raters use. Brown, Iwashita and McNamara (2005) attributed a lack of clear distinction between performances at adjoining levels to "the use of holistic assessment to provide the baseline score data, rather than more specifically focused analytic scores" (p. 104). In such a case, raters faced a difficulty in making a judgment as they attempted to balance multiple features of language in their assessment (Iwashita *et al.* 2008). As the data in this study came from a beginner level course, the findings also support the observation made in Iwashita *et al.* (2008) that the lack of clear distinction between adjacent levels was especially evident in performances of lower level students. However, even when analytical scales were used, researchers still found that a lack of distinction between levels still existed (Brown, 2006).

Raters were not told whether to assess the students individually or as a pair; however, all raters chose to assess each student individually. Consequently, each student in the pair did not always receive the same score. The scores that a particular student received sometimes varied quite substantially. Nevertheless, a general pattern emerged. More precisely, there seems to be a consensus among the raters that student A clearly outperformed student B in pair one. Except for rater 8, most raters viewed students C and D as having relatively similar level of proficiency. Unanimously, student F was perceived to be stronger than student E in pair 3.

Table 1. Summary of raters' scores

Raters	Scores					
	Pair 1		Pair 2		Pair 3	
	Student A	Student B	Student C	Student D	Student E	Student F
1	A-	C+	C+	C+	C+	B+
2	B+/A	C+	C/C+	C/C+	B	B+
3	A-	B-	C+/ B-	C+/B-	B-	B-
4	A	C+	C+	C+	B-	B
5	A	C+	C	C	B	A
6	B+	C+/B	C	C	C+	B
7	46/50 A	35/50 - 36/50 C/C+	31/50 -32.5/50 D+	33/50-34.5/50 D+	37.5/50 C+	40.5/50 B
8	A	B	C+	C+ /B	B	A
9	A	B	C+	C+	C+	A
10	8 B	6.5 D+	5 F	5 F	7 C	9 A
11	16.5/20 B	14.5/20 C	12.5/20 D+	12.5/20 D+	14/20 C	15.5/20 C+
12	A	C	C+/B	C+/B	B+	B+
13	B+	B-	B-	B-	C+/B	B+
14	A	C+/B-	B	B	B-	B+
15	B+/A	C+	B	B	B	B+
16	B	C	C+/B	C+/B	C	B
17	B	C+	B-	B-	B-	B+

Table 2. Summary of scores given by rater 7

Criteria	Scores given by rater 7					
	Pair 1		**Pair 2**		**Pair 3**	
	Student A	Student B	Student C	Student D	Student E	Student F
Communicative competence (10 points)	10	7-7.5	7	7-7.5	8	9
Language use (10 points)	9	7-7.5	6-6.5	7	7.5	8
Pronunciation (10 points)	9	7	6.5	6.5-7	7.5	8
Fluency (10 points)	9	7	6-6.5	6.5-7	7	8
Naturalness (10 points)	9	7	6	6	7.5	7.5
Total 50 points	46	35-36	31-32.5	33-34.5	37.5	40.5

2. What features did the raters take into account when assessing students' proficiency in a paired speaking task?

The coding of the raters' comments revealed five main categories, including: *task completion*, *accuracy*, *interactional management*, *naturalness*,and *fluency*.Each of these categories is discussed in the following section and illustrated with examples of extracts from the verbal protocol transcripts.

1. Task completion

Task completion encompasses the extent to which students used the language covered during the course in their conversation and how well they used them. Raters seemingly had a list of language functions, structures, expressions, and vocabulary that they expected students to incorporate into their conversations. The following sample comments focus on this aspect of task completion:

> Extract 1: All the language functions that are supposed to be used were there. There's *greeting*. There's the *introduction*. And he's asking her about *music*. So all the functions are there. (Rater 4, Pair 1, Student A).

> Extract 2: It looks like she'd prepared for the exam. She was trying to use the functions and the language covered in the course. (Rater 5, Pair 1, Student B)

Task completion extends beyond a diagnostic checklist and includes other qualitative and quantitative aspects including the amount of speech that students produced, degree of elaborations of ideas, complexity of language structures, expressions and vocabulary, relevancy of the contents or ideas to the purpose of the task, and logical organization of topics or ideas. Raters 11 and 14 shared similar views of a student's performance with regard to his vocabulary use.

> Extract 3: He's definitely better. He was using some quite advanced colloquial vocabulary. (Rater 11, Pair 1, Student A)

> Extract 4: He seemed quite comfortable. He used idiomatic expressions that marked him as someone who's very comfortable. (Rater 14, Pair 1, Student A)

Rater 15 gave quite extensive descriptions of an "average" performance illustrated by both students in pair 2, adding that these characteristics were typical of most average students in the class. Notice that task completion is tied with fluency.

> Extract 5: They could do what was required of them, but not particularly fluently. They were competent in fulfilling the basic tasks that were asked of them. But they didn't do so with particular fluency. And they didn't seem to go the extra mile, or particularly add anything that would make it stand out to be particularly good. It was almost like listening to a conversation being read out of a book. With a particular phrase they said like *Oh, that's great*. It's fine, but it's not going to get them a high grade. (Rater 15, Pair 2, Students C and D)

Raters 3 and 12 shared similar views on issues of coherence for student F.

> Extract 6: The girl on the right talked a lot, but there's no logical organization of ideas. There were a lot of repetitions of the information and topics. (Rater 3, Pair 3, Student F)

> Extract 7: This conversation makes me think that logical flow of ideas is important. When I listen to them I don't just check off the functions on the list. You can't talk about random things without any organization. (Rater 4, Pair 3, Student F)

> Extract 8: She was able to keep up a stream of sound, but it's very random. She started talking about *food* and went to *music* and back to *food* again. There were no details. She's got a stream of words coming out all the time, but not much of editing. (Rater 12, Pair 3, Student F)

On the other hand, rater 17 perceived student F quite positively, complementing on her effort to contribute to the conversation.

> Extract 9: She's put in a lot of good words, good structures, things that average students don't know. That shows that she's better. (Rater 17, Pair 3, Student F)

Most raters expressed similar views about student E, who was student F's partner from pair 3. Her contribution in the conversation was significantly less than that of her partner, and this was probably part of the reason why she received lower scores from most raters when compared to her partner.

> Extract 10: She didn't talk much. Her responses were quite short. (Rater 13, Pair 3, Student E)

Rater 2 raised the issue of relevancy of contents in pair 3, commenting that students E and F did not, to a great extent, relate the contents of their conversation to its context, which was a mutual friend's birthday party.

> Extract 11: Overall, it's fine. But I think they could have talked more about the party itself, instead of other random topics. (Rater 2, Pair 3, Students C and D)

Based on the extracts discussed above, it is evident that the raters attended to both the quantitative and qualitative aspects of task completion. Raters expected students to use certain language functions, structures, and vocabulary in their performance. This diagnostic checklist appears to be used to form a baseline for an "average" performance. In other words, students are expected to fulfill the main objectives of the task, which are to introduce themselves and then get to know each other by asking and answering questions. Good performances are distinguished from weaker ones with various factors. Similar to raters in Brown *et al.* (2005), raters in this present study commented on the amount of speech produced in terms of sufficiency for the task. However, the quantity is weighted against other qualitative aspects including relevancy and logical organization of ideas (Extracts 6, 8, 11). When judging the completeness of the task, a student's performance can be compared to that of the partner (Extract 3) and to other performances the raters have experienced (Extract 9).

2. Accuracy

Most speaking criteria make reference to the accuracy of test takers' language, and the ones used by the raters in this present study are no exception (the following extracts referenced are located below). Comments on accuracy found in this study pertain to two main linguistic features, including grammar and pronunciation. Accuracy was mainly perceived holistically based on the overall amount of errors students made as evident in the use of terms like "no errors", "some mistakes", or " a lot of mistakes" (Extracts 12, 13, 14, 15) and frequency such as "at times" (Extract 19). Accuracy was also judged in terms of severity of errors (Extracts 18, 22, 23, 24, 25).

The raters seemed to focus mainly on sentence-level grammatical accuracy (Extracts 12, 13, 14, 16, 21). Comments regarding the extended discourse such as connectives, discourse markers, and other cohesive devices as reported in Brown *et al* (2005) were not found in this study. The raters' main focus was on the correct use of language structures, especially those covered in the textbook or in class. In other words, the raters were particularly critical and attentive to inaccurate use of the language structures that students should have mastered at this stage (Extracts 16, 17). Some raters noted errors of verb forms and choice of tense (Extracts 15, 19, 22). The ability to correct oneself and correct a partner was considered an indicator of a strong test taker (Extracts 12, 14). Rater 14 in extract 20 commented on accuracy in relation to the range of language structures students used, stating that students did not make errors because of their repetitive use of a limited range of structures.

Some attention was also given to accuracy of pronunciation (Extracts 18, 19, 23, 26), which was weighed against two criteria. On one hand, raters had a benchmark by which the students' pronunciation or mispronunciation was judged, as illustrated in comments like "Her pronunciation is off at times" (Extract 19). More examples can be found in extract 26 which contained a comment "She had a strange pronunciation" and the speculation that "Maybe she wanted to sound 'farang'". (In this context, the Thai word "farang" is roughly translated as "a native speaker"). Another factor taken into consideration along with the benchmark was comprehensibility of pronunciation as illustrated in rater 7's comment in extract 18. More precisely, the rater did not expect students to have a completely accurate pronunciation as long as intelligibility of utterances did not suffer. In other words, comprehensibility rather than accuracy influenced raters' assessment of pronunciation.

> Extract 12: He made some errors, but he's able to correct himself. He corrected his partner too. (Rater 1, Pair 1, Student A)
>
> Extract 13: There were no syntactical errors. (Rater 12, Pair 1, Student A)
>
> Extract 14: He almost didn't make any grammar mistakes, and he corrected his partner too. (Rater 13, Pair 1, Student A)
>
> Extract 15: He made some mistakes with the verbs, not a lot. (Rater 17, Pair 1, Student A)
>
> Extract 16: The girl had some errors, but these didn't lead to communication breakdown. These are typical errors made by most average Thai students speaking English. (Rater 2, Pair 1, Student B)
>
> Extract 17: She hasn't mastered some basic grammar points that she should have been able to use

correctly at this stage. (Rater 5, Pair 1, Student B)

Extract 18: For pronunciation, she has a Thai accent, but I won't penalize her for that. It can be improved. But there are certain parts where she made pronunciation errors that impeded understanding. (Rater 7, Pair 1, Student B)

Extract 19: She could have done better with the verbs, some simple things that come up all the time in class like "*I like to listening music*", "*I have to dinner*"… Her pronunciation is off at times. (Rater 17, Pair 1, Student B)

Extract 20: They didn't really have grammar errors. They had a couple of simple phrases nailed down and they kept repeating those. (Rater 14, Pair 2, Students C and D)

Extract 21: The one on the left used the right tenses. She used the past tense when she's talking about the past. And she used it quite correctly too. But she just didn't talk enough. (Rater 2, Pair 3, Student E)

Extract 22: She had a lot of incomplete sentences. (Rater 6, Pair 3, Student E)

Extract 23: She had only some minor grammar mistakes, and there was no major pronunciation problem. (Rater 9, Pair 3, Student F)

Extract 24: Grammatically, she wasn't making major errors, just dropping articles, prepositions, and adding them where they didn't belong. (Rater 15, Pair 3, student F)

> Extract 25: She's comfortable with her level of English. She doesn't see the need to work on her grammar. She was making some awful grammar mistakes. (Rater 11, Pair 3, Student F)
>
> Extract 26: She's pretty fluent. She spoke very quickly, so she made a lot of mistakes. She had a strange pronunciation. Maybe she was trying to sound "farang". (Rater 16, Pair 3, Student F)

To sum up, accuracy was conceptualized in terms of the overall amount and severity of grammatical and pronunciation errors. Particular attention was given to errors of verb forms and tense uses. Interestingly, errors in other linguistic aspects of the performance such as vocabulary use or cohesive devices were not salient to the raters.

3. Interactional management

In performance-based language assessment that involves interactions between individuals, the social nature of the performance cannot be overlooked (McNamara, 1997). The co-construction of discourse by both test-takers allows for an opportunity for turn-taking, initiation of topics, and extended discourse with a partner rather than an interviewer (May 2011). A number of comments made by the raters in the present study indicate that they recognized the intrinsically co-constructed nature of the performance as illustrated in comments pertaining to turn taking, holding and yielding the floor, introducing, and building the topic.

Management of turn is vital to the students' opportunity to contribute to the conversation, which in turn affects the raters' judgment of the students' performance. The following extracts illustrate the raters' recognition of an asymmetrical pattern of interaction as a result of an unequal contribution from each interlocutor.

> Extract 27: She dominated the conversation, like there's no turn taking. She just kept asking her partner questions. (Rater 2, Pair 3, Student F)
>
> Extract 28: This [conversation] was not fair to the girl on the left [Student E]. Her partner [Student F] totally dominated a conversation. She didn't have much chance to speak, but when she did, her follow up questions were pretty good. She understood what her partner said but she didn't have much chance to speak. (Rater 8, Pair 3, Students E and F)
>
> Extract 29: She seemed quite prepared to let her partner lead. She's comfortable to say very little. She didn't have the opportunity to speak. She didn't really try. (Rater 15, Pair 3, Student E)

The asymmetrical pattern of interaction leads to two somewhat contradicting perceptions of the performance. On one hand, raters 2 and 8 viewed the domination of conversation quite negatively, and rater 8 appeared sympathetic to the other student whose chance to contribute to the conversation was perceived to be diminished by her dominating partner. On the contrary, rater 15, though recognizing that student E's chance to speak was minimized, viewed the passive student in a negative light, commenting on her willingness to let her partner take a lead in the conversation. Despite these conflicting perceptions, these raters, and in fact most of the raters awarded student F higher scores than her partner. It is possible to speculate that a passive role in a conversation is more likely to have a negative impact on scores than a dominant role is.

In addition to management of turns, the raters expected students to display abilities to initiate topics, sustain a conversation and maintain coherence in a conversation using appropriate follow-up questions. Some of the raters made reference to the functions that were covered in the course such as "showing interest", and "ending a conversation" (Extracts 31, 33).

> Extract 30: She didn't initiate a conversation much. She waited for her partner to ask her questions, and she responded to questions, especially at the beginning. (Rater 1, Pair 1, Student B)

> Extract 31: His strong point is that he asked a lot of follow up questions. And he showed interest with *really?* He made the conversation sound more interactive. He had interactional skills. (Rater 5, Pair 1, Student A)

> Extract 32: He was able to come up with comments that showed that he's capable. He was making comments at the right time, and the right place. Other students know these questions but can't use them correctly. He used them correctly. (Rater 15, Pair 1, Student A)

> Extract 33: She had follow up questions. She showed interest. There was an attempt to close a conversation. (Rater 5, Pair 3, Student F)
>
> Extract 34: She didn't ask a lot of follow up questions to keep the conversation going, just a lot of *how about you?* (Rater 12, Pair 3, Student E)

It is worth noting that comments pertaining to interaction management were mainly found in pairs 1 and 3. One similarity between these two pairs is that raters viewed a marked difference in terms of the quantity of talk, with one student leading the interaction and the other taking a more passive role. However, as illustrated in the comments, raters criticized student F for not yielding the floor, while student A was complimented for his role in moving the conversation along with his interactional skills. These different perceptions could be explained by findings from previous research studies on the impacts of conversation styles of higher-proficiency level interlocutors on their lower-proficiency level partners (See Galaczi, 2008; Lazaraton& Davis, 2008; and Nakatsuhara, 2006). In short, Galaczi (2008) pointed out that interlocutors' dominating role was perceived negatively when it was domineering rather than facilitating. Relating this finding to this present study, student F might have been viewed as a domineering partner rather than a facilitating one like student A. Stronger interlocutors in Nakatsuhara's (2006) study displayed similar conversational accommodation that helped make the conversation appear more conversational and collaborative. These facilitative or accommodating behaviors can also be construed as identity construction. More precisely, Lazaraton and Davis (2008) asserted that when stronger test takers were paired with weaker partners, one way to construct and reaffirm their identity as proficient speakers who were deserving of high scores on the test was to take on a supportive role in the conversation. This role was reflected in the various ways they scaffold their weaker partners such as expanding on partners' contributions or handling partner's problematic talk. In relation to this present study, raters' comments seemed to indicate that student A successfully cast himself as a proficient interlocutor through using appropriate follow-up questions, as well as correcting his partner's language (as illustrated in extract 14 in the section about *Accuracy*).

Evidently, interaction management becomes a salient issue for raters when asymmetrical patterns of interaction emerge in the performance. Previous research on interlocutor effect has raised a concern regarding fairness of scores awarded to each test taker. Rater 8's comment in extract 28 suggests that the rater recognized the negative impact of one test taker's performance on the other test taker's discourse. However, unlike May's (2009) study, there was no evidence in the comments to suggest that the rater considered penalizing or compensating for the test taker's role in the interaction.

4. Naturalness

For the raters in this study, naturalnessencompasses characteristics of a conversation and those of the interlocutors that resemble what may be found in an authentic conversation. The comments that fall in this category include the raters' overall assessment of the performance, test takers' demeanor, personality, as well as other kinds of non-verbal communication such as gestures or eye gaze.

Positive comments on student A's performance mostly include non-verbal features that characterize a natural interaction (see extracts below). The raters referred to student A's demeanor with descriptive words like "natural", "relaxed", "comfortable", "confident", "in control", and "cheeky". Rater 15 considered student A's ability to joke around as an indicator of the level of comfort and confidence that the student might have had. In addition to demeanor, another non-verbal feature that was salient to rater 10 was eye contact (Extract 35). The use of verbs like "sounded" and "appeared" in delivering their comments seems to indicate that the raters observed these features within the speaker's performance.

> Extract 35: He sounded more natural. He had better eye contact. His position was more dominating, but relaxed. (Rater 10, Pair 1, Student A)
>
> Extract 36: He appeared very comfortable with the language. He's confident (Rater 12, Pair 1, Student A)
>
> Extract 37: He was in control of what he was saying. He was able to throw in a little humor, a little cheeky to the girl. He's confident. (Rater 15, Pair 1, Student A).

A completely different picture is painted by comments for both students in pair 2. Their interaction was mostly seen as lacking naturalness and characterized as “memorized”, “mechanical”, “formulaic”, “stilted”, “awkward”, “not natural”, “rehearsed”, and “robotic”. There were references to the display of discomfort (Extracts 42, 44). Raters also mentioned ineffectiveness of the co-construction of discourse with comments like, “They were prompting each other” (Extract 39), and “They were trying to follow a list” (Extract 43). One rater also observed a lack of good eye contact that could be inferred from the comments in extract 44.

> Extract 38: The whole thing sounds very memorized. (Rater 6, Pair 2, Students C and D)
>
> Extract 39: They were very mechanical. They were prompting each other. (Rater 10, Pair 2, Student C and D)
>
> Extract 40: The conversation sounded very formulaic. (Rater 15, Pair, 2, Students C and D)
>
> Extract 41: It was a stilted conversation, from stock phrases that they have remembered. (Rater 11, Pair 2, Students C and D)
>
> Extract 42: They didn’t look enormously comfortable. They sounded memorized (Rater 14, Pair 2, Students C and D)
>
> Extract 43: The conversation sounds awkward. It’s not natural. They were trying to follow a list. (Rater 17, Pair 2, Students C and D)

> Extract 44: She looked up a lot, probably thinking about what to say. It seemed rehearsed and robotic. She could have been more relaxed. (Rater 10, Pair 1, Student B)

The following section presents the raters' comments for pair 3, mostly for student F. Features that characterized naturalness were a display of emotion, and the use of varying tones (Extract 46). According to rater 15, these features contributed to a "genuine" conversation. The raters made references to student F's personality with words like "outgoing", "enthusiastic", and "confident" and a display of comfort and a positive attitude. Rater 17 in extract 48 made an overall assessment of pair 3's conversation, which suggested a comparison with the use of the comparative "more realistic". Given that the raters viewed pair 3's conversation after that of pair 2, it was likely that the comparison was in reference to that of pair 2.

> Extract 45: She seemed outgoing. (Rater 10, Pair 3, Student F)

> Extract 46: She's very enthusiastic and confident in her abilities to do everything. One thing that she was doing that makes it stand out from the middle pair is that she was using tones a lot more to show actual genuine, um not genuine emotion, but trying to show some emotion. She gave the appearance of a more genuine conversation by changing the tone from time to time. Maybe it's a personality thing. (Rater 15, Pair 3, Student F)

Extract 47: She seems comfortable and natural. She has positive attitude. (Rater 17, Pair 3, Student F)

Extract 48: This conversation sounds more realistic. (Rater 17, Pair 3, Students C and D)

Overall, the comments pertaining to naturalness of the performance were benchmarked against the raters' concepts of a natural conversation. Although the descriptions they used may appear intuitive, they seem to be based on a common knowledge of certain characteristics of a natural conversation. It is important to note that the raters' comments in this section are directed either toward an individual test taker or toward both test takers in each pair. In this respect, May (2011) discussed whether test takers' contribution in a co-construction interaction was separable. She found that certain features of interaction were likely to be perceived as mutual accomplishment, one of which was "contributes to an authentic interaction" (p. 139). For the most part, the findings presented in this section concur with May's. However, it is possible to further conclude that when there was no marked difference between the performance of each student in the pair (as evident in most of the comments for pair 2), the raters would more likely assess the performance as mutual achievement.

5. Fluency

In Fulcher's (1996) study, one of the most extensive explorations of the construct "fluency", the researcher identified eight aspects of performance that accounted for observed interruption of fluency. These aspects included:

1) End-of-turn pauses: pauses indicating the end of a turn.
2) Content planning hesitation: pauses which appear to allow for the student to plan the content of the next utterance.
3) Grammatical planning hesitation: pauses which appear to allow the student to plan the form of the next utterance.
4) Addition of examples, counterexamples, or reasons to support a point of view: these pauses are used as an oral parenthesis before adding extra information to an argument or point of view, or break up a list of examples.
5) Expressing lexical uncertainty: pauses which mark searching for a word or expression.
6) Grammatical and/or lexical repair: hesitation phenomena which appear to be associated with self-correction.
7) Expressing propositional uncertainty: hesitation phenomena which appear to mark uncertainty in the views which are being expressed.
8) Misunderstanding or breakdown in communication (p. 217).

The raters in this study also associated features such as pauses and hesitation to a lack of fluency; however, these comments did not always provide explanations or speculations for interruption of fluency like those provided in Fulcher's categories. The majority of comments on fluency were an overall assessment expressed in broad terms suggesting varying degrees of fluency such as "fluent" "quite fluent", and "not fluent".

> Extract 49: He's fluent. He didn't seem to struggle when he initiated a topic. (Rater 2, Pair 1, Student A)
>
> Extract 50: He didn't stumble at all. (Rater 12, Pair 1, Student A)
>
> Extract 51: She's not fluent at all. (Rater 16, Pair 1, Student B)
>
> Extract 52: They could do what was required of them, but not particularly fluently. They were competent in fulfilling the basic tasks that were asked of them. But they didn't do so with particular fluency. (Rater 15, Pair 2, Students C and D)

More specific comments were also found in the data, but the raters often referred to features associated with non-fluency such as pauses, and hesitation. Extracts 53 and 54 focus on the length of pauses that negatively affected the flow of a conversation. One rater also compared the degree of fluency between the two test takers (Extract 55). In extract 56, rater 12 compared a fluent speech to a fluent execution of piano scales. This comment seems to suggest that fluency was also conceptualized in terms of speech rate.

> Extract 53: They took a lot of time to think before they spoke. They couldn't keep the conversation going very smoothly. (Rater 3, Pair 2, Students C and D)
>
> Extract 54: There were a lot of long pauses. She took a long time before she responded. Sometimes it was so long that as a listener, it was uncomfortable,

like what was she trying to say? (Rater 7, Pair 1, Student B)

Extract 55: The boy seemed quicker with answers and more ready to start a conversation. He's better at thinking on his feet. (Rater 15, Pair 2, Student D)

Extract 56: They hesitated. They just need more practice. It's like when you practice the scales on the piano. They got all the notes right, but they're doing the scales very slowly. (Rater 12, Pair 2)

Fluency was sometimes measured against other aspects of the performance including accuracy (Extracts 57, 58) and logical organization of ideas (Extract 59). In the raters' perspectives, student F may have concentrated on fluency at the expense of accuracy and coherence.

Extract 57: She might appear fluent, but she had a lot of errors (Rater 2, Pair 3, Student F)

Extract 58: She's pretty fluent. She spoke very quickly, so she made a lot of mistakes. (Rater 16, Pair 3, Student F)

Extract 59: She's got a stream of words coming out all the time, but not much of editing. She confuses fluency with streams of words. (Rater 12, Pair 3, Student F)

Summary and implications of findings

This study originated from the need to understand the approaches that different course instructors use in assessing a paired speaking task when no specific criteria or rating scales are provided. The findings indicate that most of the raters, with one exception, used holistic scoring approach when assessing students' performance, assigning either one single letter grade or a numerical score to each student in the pair. The raters were asked to discuss various features of the conversation that they attended to when assessing the students' performance. The results of verbal protocol analysis formed five main features, including "task completion", "accuracy", "interaction management", "naturalness", and "fluency".

An assessment of speaking proficiency is a task plagued by subjectivity, even with assessment criteria. Without any criteria, raters are left to their own devices, relying only on their judgment and intuitions. While the aim of this study was not to construct rating scales, the findings in the present study indicate possible common criteria that are deemed important by raters when assessing a paired test task at a beginner level. As such, these criteria contribute to the first step in constructing rating scales for similar types of task in Thai contexts. Raters' comments in verbal protocols can be used as descriptors in the rating scales. This is particularly helpful in a classroom assessment context as the teacher-raters can more easily justify awarded scores, and at the same time allow students to learn about their strengths and weaknesses, giving them more specific goals to work toward improving their performance. Also, for a compulsory or even elective course that is offered to a large numbers of students, hence requiring multiple instructors to teach different groups of students, having a shared understanding among all instructors as to how students are to be assessed will likely lessen a concern about fairness.

The results also suggest that the very feature that makes a paired test attractive as a test task can complicate assessment. In particular, the co-constructed nature of the task means that one test taker's performance will likely impact the other's performance either positively or negatively to a certain degree. The success of the interaction is a shared responsibility among the interlocutors, and as such "interactional competence" should not be considered an attribute of individual interlocutors (Jacoby & Ochs, 1995; Kramsch, 1986; McNamara, 1997). These influences were observed by the raters and expressed in their comments in the "interaction management" section. As the data in this study suggested, such influences appear most problematic to the raters when the conversation is characterized as asymmetric with one test taker dominating the conversation in ways that disadvantage rather than accommodate the other. To deal with this potentially problematic issue, the shared nature of interactional competence should be included in the descriptors of the rating scales whether they are holistic or analytic. If analytical rating scales are used, a shared score for interactional management can be awarded to both interlocutors.

Limitations and future work

Given the small number of raters participating in this study, the results cannot be generalized in other testing contexts. However, the results serve as a solid starting point for the development of empirically based rating scales for a specific testing context. The next step will be to allow course instructors to try implementing the scales in their real testing context to determine if other constructs should be added and whether the descriptors can be enriched and revised. Another important area worth further investigation is to map particular features of the interaction that the raters associated with different strong or weak performances. Such analysis would further improve the rating scales that are effective in distinguishing the various levels of performance. After all, that is what any test is meant to achieve.

References

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

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.), *IELTS research reports 2006* (pp. 71-89). Canberra & Manchester: IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for Academic-Purposes speaking tasks.* (TOEFL Monograph No. 29). Princeton, NJ: Education Testing Service.

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

Douglas, D., & Smith, J. (1997). Theoretical underpinnings of the Test of Spoken English revision project. *TOEFL Monograph Series 9*. Retrieved from http://www.ets.org/research/policy_ research_ reports/rm-97-02_toefl-ms-09.

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

Együd, G., & Glover, P. (2001). Oral testing in pairs-a secondary school perspective. *ELT Journal, 55*(1), 70-76.

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

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Galaczi, E.D. (2008). Peer-peer interaction in a speaking test: The case of the First Certification in English. *Language Assessment Quarterly, 5*(2), 89-119.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher and lower-scoring students. *Language Testing, 27*(4*),* 585-602.

Glaser, B. G., & Strauss, A.L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine de Gruyter.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook Vol. 5.* Cambridge: Cambridge University Press.

Ikeda, K. (1998). The paired learner interview: A preliminary investigation applying Vygotskian insights. *Language, Culture and Curriculum, 11*(1), 71-96.

Iwashita, N. (1996).The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-66.

Iwashita, N. (1997). Assessment of oral communication skills in LOTE settings in Australia.*MelbournePapers in Language Testing, 6*(2), 37-43.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Jacoby, S., & Ochs, E. (1995). Co-construction: an introduction. *Research on Language and Social Interaction, 28*(3), 215-231.

Kormos, J. (1999). Simulating conversation in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing 16*(2), 163-188.

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

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

Luoma, S. (2004). *Assessing Speaking*. Cambridge, UK: Cambridge University Press.

Macqueen, S., & Harding, L. (2009). Review of the Certificate of Proficiency in English (CPE) speaking test. *Language Testing, 26*(3), 467-475.

McNamara, T. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

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

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

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

Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired speaking tests. *Research Notes, 25*. Retrieved from www.cambridgeesol.org/rs_notes/rs_nts25.pdf

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

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

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186.

Orr. M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System 30*, 143-154.

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

Pollitt, A., & Murray, N. (1996). 'What raters really pay attention to.' In M. Milanovic, & N. Saville (Eds.), *Performance Testing, Cognition and Assessment: Selected Papers from the 15th. Language Testing Research Colloquium (LTRC)* - Cambridge and Arnhem 1993 (pp. 74-91). Cambridge: Cambridge University Press.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal 53*(1), 42-51.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3)*,* 275-302.

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

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*(3)*,* 325-339.

UCLES. (2008a). Cambridge English Key English Test (KET): Handbook for Teachers. Retrieved from https://www.teachers.cambridgeesol.org/ts/digitalAssets/117391_Cambridge_ English_Key__KET__Handbook.pdf

UCLES. (2008b). Certificate of Proficiency in English: Handbook for Teachers. Retrieved from https://www.teachers.cambridgeesol.org/ts/digitalAssets/117377_Cambridge_English_Proficiency__CPE__Handbook.pdf