Exploring shared and individual assessment of paired oral interactions

Pakize Uludag, Kim McDonough & Pavel Trofimovich
Concordia University

Studies concerning the assessment of second language (L2) paired oral interaction to date have investigated interactional patterns that emerge from paired oral tests and identified the factors that create variability in an individual's test scores. However, less work has addressed concerns about whether scores should be shared or individual when assessing L2 speakers' oral interactions. Therefore, the present study compared shared and individual assessment of L2 English paired oral task performances. Paired oral interaction episodes were sampled from a larger corpus of university-level L2 speakers engaged in paired speaking tasks and were assessed by 60 raters who were randomly assigned to rate Speaker A, Speaker B, or both speakers. To avoid any possible rating effects due to rating stimuli, half the raters evaluated audio recordings while the other half assessed video recordings. The raters used an analytic rubric with four domains: discourse management, collaborative communication, content development, and language accuracy and complexity. Comparison of raters' scores revealed that individual discourse management ratings were significantly higher than shared ratings for both members of the pair regardless of the rating modality (audio vs. video). Implications for assessing pair interactions are discussed.

Key words*:* paired speaking tasks, language testing, individual and shared assessment

Exploring shared and individual assessment of paired oral interactions

To reflect authentic communicative language use more closely, there has been a shift from examiner–candidate oral testing to assessment through paired speaking tests in second

Email address for correspondence: pakize.uludag@concordia.ca

language (L2) assessment (Norton, 2005, 2013). In contrast to examiner–candidate interviews, where an examiner plays a neutral role and leads the interaction, paired speaking tests elicit co-constructed performance between two speakers of more equal standing through collaborative interaction (Lazaraton, & Davis, 2008; Taylor, 2001; Vo, 2019). This format has gained in popularity due to concerns about the power relationship between an examiner and examinee in traditional interview-style L2 proficiency tests, such as the Oral Proficiency Interview (OPI) and Stimulated Oral Proficiency Interview (SOPI). Although researchers focusing on the asymmetric nature of examiner–candidate interviews have widely discussed the advantages of paired speaking tests, questions have been raised about whether test performance should be evaluated as an individual or joint achievement (e.g., O'Sullivan; 2002; Taylor & Wigglesworth, 2009). In addition, validation studies of paired-speaking tests have found that nonverbal aspects of communication were salient to the raters (e.g., Ducasse & Brown, 2009; May, 2011; Orr, 2002), with few studies examining paired interaction ratings based on audio and video recordings (Beltrán, 2016; Nambiar & Goon, 1993; Styles, 1993). Therefore, this study compared shared and individual ratings of paired oral interactions through the use of video and audio recordings.

Paired speaking tests

Reflecting the importance of interaction in L2 learning and teaching, the paired test format has been increasingly used for placement, exit, and achievement test purposes at language schools and universities (Brooks, 2009; Ducasse & Brown, 2009). Unlike examiner-led tasks, where the interviewer asks questions and the test taker responds, paired speaking tasks elicit a more balanced distribution of turns and a greater range of interactive features and functions (Brooks, 2009; Kormos, 1999; Lazaraton, 2002). Analysis of paired speaking task interaction has indicated that language features produced by test takers tend to be equally distributed (i.e., symmetric) since one speaker does not dominate the other (Brooks, 2009; Taylor, 2001). Furthermore, because pair and group work activities are commonly used in communicative language classrooms, using a paired format to evaluate L2 speakers' oral proficiency generates positive washback as test takers experience less anxiety while interacting with a peer (Együd & Glover, 2001; Galaczi, 2008). As such, paired speaking tasks most likely elicit more authentic conversations compared to interviewer-led interaction, which has greater similarity to institutional talk (van Lier, 1989).

Researchers exploring the collaborative nature of pair interactions have shown that a variety of factors create variability in an individual's test scores, including proficiency,

personality, and interlocutor familiarity (Galaczi, 2008; Nakatsuhara, 2011; O'Sullivan, 2002), which has raised questions about the validity of paired speaking tests. For example, drawing upon qualitative analysis, Norton (2005) investigated the impact of interlocutor proficiency on speakers' task performance during the Cambridge speaking tests. The results indicated that low proficiency test takers paired with higher proficiency candidates benefited from the language produced by their conversation partners. In the case of test taker personality, Berry (2007) reported an interaction between personality and task type, with introverts receiving higher scores when paired with partners with the same personality type. Focusing on interlocutor familiarity, gender, and language proficiency, O'Sullivan (2002) found an interaction between acquaintanceship, cultural belonging, and gender. More specifically, L2 speakers paired with a friend performed better than when they were paired with a stranger. Taken together, these studies have identified sources of variability in the assessment of an individual's task performance on paired speaking tests.

Rating focus in paired speaking tests

Since the discourse is co-constructed through the collaborative effort of both interlocutors during pair/group interactions, L2 speakers' test performances are inextricably linked (Luoma, 2004; McNamara, 1997). Thus, researchers have debated whether paired or group oral assessment is best interpreted as an individual or shared achievement and how individual scores from jointly constructed interaction should be interpreted (May, 2011; O'Sullivan; 2002; Swain, 2001; Taylor & Wigglesworth, 2009). For example, May (2011) found that raters interpreted key features of the paired test interaction, such as understanding and responding to their partner, working collaboratively, and contributing to the quality of the interaction, as a mutual achievement rather than an attribute of an individual speaker.

Acknowledging the interdependence between interlocutors' contributions to conversation, some assessment researchers have argued that L2 performance involving interactions should be evaluated in terms of symmetrical and collaborative turn-taking behavior and topic management (e.g., Chalhoub- Deville, 2003; McNamara, 1997). In this approach to shared ratings, performance is evaluated in terms of how well the interlocutors collaborate by managing various discourse features, which are typically captured in research through such interactional measures as length and order of turns, their distribution, and topic nomination (e.g., Lazaraton & Davis; 2008; Nakatsuhara, 2011). For example, using methodological tools from conversation analysis, Galaczi (2004) investigated 30 test taker dyads from the Cambridge First Certificate in English and identified four distinctive interactional features: collaborative, parallel, asymmetric, and blended. The majority of the test takers used collaborative, parallel, or blended

patterns of interaction, each representing 30% of the dataset, while asymmetric interactions (one dominant and one passive interlocutor) did not occur frequently (10% of the dataset). In addition, test takers using a collaborative pattern of interaction were found to obtain the highest scores on the interactive communication scale, indicating that raters can evaluate discourse features across different score bands.

However, when awarding shared scores for interactional competence for paired tests, raters might be unduly influenced by one interlocutor's contribution to the conversation (Brown, 2003; Davis, 2009; Galaczi & Taylor, 2018). In other words, an L2 speaker participating in interaction might receive a lower (or higher) shared score than they would have received if scored individually. To explore this possibility, McDonough and Uludag (2021) investigated potential differences in shared and individual ratings of paired speaking tests. Using an analytic rubric with four criteria (discourse management, collaboration, content development, and language accuracy), four raters evaluated each pair's shared performance while two raters evaluated an individual speaker's performance. Comparison of individual and shared ratings across multiple speaker pairs revealed that shared discourse management ratings were significantly higher than the individual ratings. However, there were no significant differences in the ratings received by each test taker in a pair. Although these results provide support for both individual and shared assessment, further studies are needed to refine the construct of interactional competence in paired and group speaking tests and provide insight into shared versus individual score interpretation (May, 2011; Swain, 2001).

In addition to debate about shared versus individual assessment, researchers have also raised questions about whether having access to visual input (i.e., nonverbal aspects of communication) impacts raters' assessment of paired or group task performance (e.g., Jenkins & Parra, 2003; May 2009, Nakatsuhara et al., 2021; Orr, 2002). To capture the impact of rating stimuli in task assessment, a few studies have compared scores from raters who evaluated either audio or video stimuli. For example, Beltrán (2016) found no difference in the ratings of fluency, pronunciation, vocabulary, grammar, and meaningfulness assigned to L2 speakers' performance on a monologic task by raters who used either audio or video recordings. An earlier International English Language Testing System (IELTS) study by Styles (1993) investigated three examiners' post hoc inter- and intra-rater correlations and reported higher audio ratings than video ratings. In contrast, research by Nambiar and Goon (1993) compared ratings of fluency, accuracy, effectiveness, and range in two different tasks (i.e., interviews and paired oral tasks) by raters who either rated the face-to-face interaction in real time or subsequently rated performance based on audio recordings only. They found that the ratings of the audio recordings were significantly lower than those given to the face-to-face interaction in both

tasks. Similar results have been reported by Nakatsuhara et al., (2021) in a recent study which compared IELTS examiners' ratings of live, audio-, and video-recorded performances and found significantly lower audio ratings than live and video ratings. Taken together, these inconsistent findings suggest possible rating effects due to rating stimuli and provide compelling evidence for the use of audio and video recordings when assessing paired performance.

In summary, questions remain as to best practices in the assessment of paired oral interactions, specifically whether scores should be shared or individual (McDonough & Uludag, 2021; O'Sullivan; 2002; Taylor & Wigglesworth, 2009). Additional debate has concerned rating stimuli, with prior studies reporting conflicting findings for the use of audio or video recordings for the assessment of paired interactions (Beltrán, 2016; Nambiar & Goon, 1993). Therefore, to shed further light on these issues, this study compared shared and individual assessment of L2 paired oral interactions using either video or audio recordings as the rating stimuli. The research question was as follows: Is there a difference in shared and individual ratings of English L2 speakers' paired oral interactions when raters evaluate either audio or video recordings?

Method

Paired oral interactions

Six, 10-minute paired oral interactions were sampled from The Corpus of English as a Lingua Franca Interaction (CELFI), which is a collection of dyadic interactions of L2 English speakers at English-medium universities in Canada (McDonough & Trofimovich, 2019). As degree-seeking students, they had met the minimum English proficiency required for admission to their universities (minimum TOEFL iBT score of 75 or equivalent). They were randomly assigned to pairs to interact with someone from a different language background, and there was an equal distribution of pairs with same and different reported genders. The CELFI involved three 10-minute interactive tasks that required speakers to engage in discussions. This study sampled L2 speaker interaction during the first task, in which they discussed challenges that international students face when moving to Quebec and suggested solutions. The task was introduced to the speakers with a 2-3 min warm-up, having them share the difficulties they personally experienced. The task instructions were then explained to the participants and presented on a handout, prompting them to exchange their opinions and engage in a discussion for 10 minutes.

As shown in prior studies, L2 proficiency, gender, and familiarity may impact both the quality and quantity of talk produced in paired oral interactions (Berry, 2007; Iwashita, 1998; Nakatsuhara, 2011; Norton, 2005; O'Sullivan, 2002). Thus, the sample was selected according to the following criteria: (a) identical L2 proficiency based on self-reported total standardized test scores from the Test of English as a Foreign Language (TOEFL) or IELTS, (b) a variety of first language (L1) backgrounds, (c) mixed gender pairs, and (d) no interlocutor familiarity prior to data collection. The selected samples came from six male and six female participants who were speakers of Mandarin Chinese (3), Arabic (2), Farsi (2), Spanish (2), Turkish (1), Tamil (1), and French (1). They ranged in age between 22 and 32 (*SD* = 4). The participants' self-reported standardized test scores ranged between 80 and 93 (IELTS 6.5). They had studied English for an average of 11.30 years (*SD* = 4.06) and had been living in Canada for an average of 2.8 years (*SD* = 3.29). The gender and L1 background of each pair is provided in Table 1.

Table 1. Speakers' gender and L1 background by pair

Pair	Speaker	L1	Gender
1	A	Mandarin	Male
	B	Spanish	Female
2	A	Turkish	Male
	B	Mandarin	Female
3	A	Arabic	Male
	B	Farsi	Female
4	A	Spanish	Female
	B	Tamil	Male
5	A	Mandarin	Female
	B	Arabic	Male
6	A	French	Female
	B	Farsi	Male

The audio and video recordings ranged between 9.5–10.7 minutes in length after trimming initial hesitations and dysfluencies. The videos showed both speakers' upper body (face, hand and arms, and torso) facing each other (see Figure 1). For the audio stimuli, a static image taken from the videos was shown in the interface where the raters listened to the audios. The volumes of audios and videos were normalized using MP3Gain Express 2.4.0.

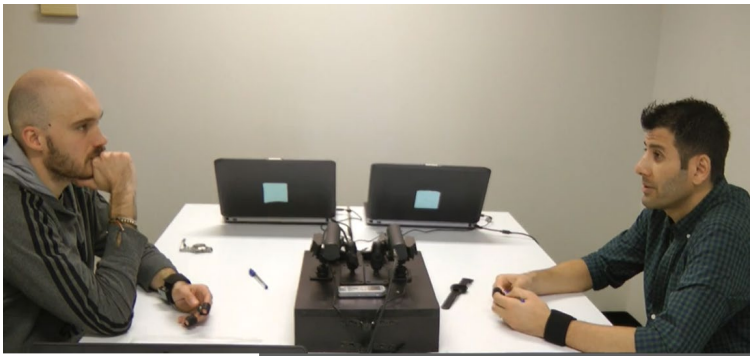

Figure 1. Screenshot of the video recordings

Rating materials

An analytic rubric (see Appendix A) was adapted from an earlier research study by McDonough and Uludag (2021), consisting of four broad domains excluding nonverbal features: discourse management, collaborative communication, content development, and language accuracy and complexity. The original rubric was adapted by including fluency in the language category and adding rating scales to each descriptor within the four domains. Discourse management was described as fluidity and organization of speech, and the use of cohesive devices based on the First Certificate of English (FCE) speaking exam criteria (Cambridge English Language Assessment). Collaborative communication was operationalized as equality (i.e., making equal contributions to the task) and mutuality (i.e., initiating, engaging with, and responding to each other's ideas) following the models of dyadic interaction (Galaczi, 2008; Storch, 2002). The content development category evaluated accuracy, relevance, and innovativeness of ideas, as well as informative reasoning through reference to Kumpulainen and Mutanen's (1999) framework. The final category, language, followed the IELTS speaking rubric descriptors, which include vocabulary range along with the appropriateness, accuracy, and complexity of grammatical structures, along with fluency.

The reliability of the original rubric had been confirmed in McDonough and Uludag (2021) through acceptable levels of interclass correlation (over .73 for shared and individual performance) as well as raters' qualitative feedback on the appropriateness of the criteria. Before using the rubrics in the current study, rater reliability and rubric category statistics were calculated implementing a Many-Facets Rasch Model using FACETS software (version 3.71.4) using shared ratings from McDonough and Uludag (2021), which included four raters. The results suggested the raters were consistent in their use of the rating scales, with infit mean square values ranging from 0.7 to 1.40 for the shared performance (Bond & Fox, 2015). As for the rubric category statistics, the average measures showed a steady increase, as do the threshold values. The outfit mean-

square values were l.0, meaning that rubric categories functioned well and that the scores were assigned in a consistent manner (Barkaoui, 2013; Linacre, 2004).

After establishing the reliability of the original rubric, slight modifications were made by giving rating scales to each descriptor in the four domains (i.e., instead of one scale per domain) and adding a fluency descriptor to the language domain. To create a rubric for assessing shared performance, wording was changed, such as changing *the speaker* to *the speakers*. The reformatted rubric was pilot tested by five research assistants, which led to further improvement of descriptor clarity (see Appendix A for both shared and individual rubrics).

Raters

The raters were 60 students enrolled in graduate and undergraduate degree programs in education departments at two English-medium universities in Canada. All but two raters had been educated entirely in English and either held or were completing degrees in applied linguistics or teaching English as a second language (TESL). They reported having English teaching experience for an average of 2.8 years (SD = 3.0), which is typical for graduate and undergraduate students studying in applied linguistics or TESL programs in this study context, except for three undergraduate raters who had never taught before. They reported using English for daily communication for both speaking (M = 81.1%) and listening (M = 84.6%). They also reported extensive exposure to L2 English and some proficiency in another language, such as French, Chinese, and Spanish. Each rater was randomly assigned to evaluate either Speaker A, Speaker B, or shared performance, which resulted in 20 raters per condition. Within the three conditions, half of the raters were randomly assigned to evaluate audio recordings while the other half assessed video recordings. Raters were assigned to only one rating group, which resulted in 10 raters in each combination of rating focus and rating modality. The background information of the 10 raters in each rating group is summarized in Table 2.

Table 2. Rater background by rating condition

Rating focus	Rating stimuli	Gender	Mean age	L1s other than English
Speaker A	Audio (n = 10)	6 women, 4 men	25.5	French (2), Spanish (1)
	Video (n = 10)	8 women, 2 men	26.4	Japanese (1)
Speaker B	Audio (n = 10)	6 women, 4 men	26.3	Farsi (1)
	Video (n = 10)	5 women, 5 men	29.0	Farsi (1), French (1)
Shared	Audio (n = 10)	8 women, 2 men	25.1	Chinese (1)
	Video (n = 10)	6 women, 4 men	28.3	Chinese (1), German (1)

Rating procedure

The raters participated in an individual rubric training session (30 minutes) with the first researcher. The training session included a review and discussion of either individual or shared rubric categories depending on the rating focus. The researcher introduced the task instructions to the raters using a practice recording (audio or video) of the same task from a pair whose performance was not included in the target materials to train the raters. After asking any questions, the raters worked independently to evaluate an additional recording and assigning a score for each rubric category.

Following the training, the rater worked independently using a personal computer with a headset in a quiet research lab (60 minutes). The raters logged into an open-access online survey interface (LimeSurvey), which presented the audios/videos in a randomized order and required them to listen to/watch the entire conversation once. Both videos and audios played automatically for each stimulus, and the replay buttons were disabled. The rubric criteria were presented to the raters on the same page as the audios/videos so that raters could assign their scores simultaneously as they were listening to/watching the interactions. The researcher (first author) remained in the office in case the raters had any questions or technical difficulties. After completing the rating task, the raters filled out a background information questionnaire eliciting information about their teaching and language assessment background (15 minutes).

To check the internal consistency of each domain in the rubric, Cronbach's alpha values were obtained. The consistency ranged between .74 and .89 for all categories in both audio and video conditions, exceeding the suggested threshold values of .70–.80 (Larson-Hall & Plonsky, 2015). Next, the subscores within each domain were summed and interrater reliability was calculated using two-way mixed intraclass correlations, which were .91 for shared performance, .95 for Speaker A, and .88 for Speaker B in the audio condition. As for the video condition, intraclass correlations were .94 for shared performance, .89 for Speaker A, and .87 for Speaker B. The ratings within each condition were averaged to derive single mean scores for shared and individual performance in the audio and video conditions.*1

Results

The research question asked about the difference in shared and individual ratings of English L2 speakers' paired oral interactions with raters using either audio or video recordings as rating stimuli. Because the overall trends in the data were similar for video and audio conditions, we present findings based on the combined data (see Appendix B

for the comparison of ratings by rating stimuli). Descriptively, except for language accuracy and complexity, the mean ratings received by either individual speaker were higher than the shared ratings, as shown in Table 3. In addition, Speaker A and Speaker B received fairly similar ratings despite some variation within the categories. Before running parametric statistics, the data were tested for the assumptions of normality and homogeneity of variances. Normality was confirmed through a Shapiro-Wilk's test, which was non-significant ($p > .05$). The assumption of homogeneity was also met as evidenced by a non-significant Levene's F test ($p > .05$). There were no significant outliers.

Table 3. Rubric scores by rating focus

Domain	Shared	Speaker A	Speaker B
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
Discourse management	5.1 (.90)	6.0 (.74)	5.9 (.90)
Collaborative communication	5.5 (.93)	6.0 (.69)	5.8 (.88)
Content development	5.5 (.96)	6.1 (.79)	6.1 (.74)
Language accuracy and complexity	8.2 (1.2)	8.2 (1.3)	7.9 (1.2)

To compare the effect of individual versus shared evaluation, four one-way ANOVAs were conducted for each rubric category. Analysis of between-group differences in rating focus yielded only one significant F-ratio—for discourse management ratings, $F(2, 57) = 6.83$, $p = .002$, $\omega = 0.60$. Bonferroni-corrected post hoc tests revealed that discourse management ratings received by Speaker A were significantly higher than both speakers' shared performance ($p = .003$, $d = 2.08$). Similarly, Speaker B received significantly higher discourse management ratings than both speakers together ($p = .024$, $d = 1.20$). However, there was no significant difference between the ratings of Speaker A and Speaker B ($p = .524$, $d = 0.24$). Turning to the remaining rubric domains, one-way ANOVAs did not reveal any significant effect of rating focus for the scores of collaboration, $F(2, 57) = 1.30$, $p = .280$, content development, $F(2, 57) = 3.23$, $p = .060$, or language accuracy and complexity $F(2, 57) = .291$, $p = .749$.

Discussion

This study compared shared and individual ratings of L2 paired oral interactions in four rubric domains when raters evaluated either audio or video recordings. To summarize the main findings, the focus of assessment appeared to make a difference for the rating of discourse management, where individual ratings were significantly higher than shared

ratings for both members of the pair. By contrast, the shared and individual ratings for collaborative communication, content development, and language accuracy and complexity were similar.

With respect to discourse management, as evidenced by large effect sizes (d = > 1.0), both Speaker A and Speaker B received significantly higher ratings compared to shared performance. On the one hand, language testers, including McNamara (1997) and Luoma (2004), have previously acknowledged that the responsibility for the creation of discourse (i.e., organization of ideas and consistency of speech) during paired and group interactions may be shared between the speakers. In addition, some features of interaction inherent in co-created discourse might be interpreted by raters as mutual achievement rather than an individual accomplishment (May, 2009, 2011). On the other hand, the obtained difference between individual and shared discourse management ratings in the present study indicates that our raters likely considered L2 speakers' performances singly rather than jointly.

One possible explanation for this finding could be that the raters anticipated a greater range of interactive features and functions from a jointly constructed discourse than individual speaker performances (Brooks, 2009; Kormos, 1999; Lazaraton, 2002). Additionally, the raters might have considered speakers' turn taking as a discourse feature, such that they factored idea sharing and reciprocal feedback in their individual ratings of each speaker's discourse management. Given that organization of turns has been recognized by raters as contributing to successful interaction in previous research (e.g., Ducasse & Brown, 2009; ffrench, 1999), an important priority for future research is to identify which key aspects of discourse management are attended to by raters when making judgements of individual and shared performances.

Our findings for discourse management were contrary to those reported by McDonough and Uludag (2021), in which raters assigned higher discourse management ratings for the shared performances in comparison to individual speaker performances. The directional discrepancy between the present research and McDonough and Uludag (2021) might pertain to the variation in task conditions and task type differences. The pairs in the current research were prompted to carry out a 10-minute discussion during which individual speakers had ample time to revisit their ideas and rephrase their opinions using discourse features. In McDonough and Uludag (2021), on the other hand, the decision-making tasks were administered under time pressure, where each pair had 2 minutes to plan and 3 minutes to talk. Possibly, the length of the conversations challenged the raters to recognize individual contributions to discourse management, leading them to award higher discourse management rating for the shared performance. Therefore, it

may be argued that L2 speakers could attain more successful peer-to-peer interaction and contribute a wider range of discourse features during prolonged interactions.

Apart from discourse management, we found no statistical difference in individual and shared ratings for collaborative communication, content development, or language accuracy and complexity. This finding is consistent with McDonough and Uludag (2021), who also reported that raters assigned identical scores for individual and shared performance for these rubric domains. Although research comparing individual and joint performance drawing on rater judgements is scant, collaborative interactions have been associated with higher scores in previous work conducted within the framework of conversation analysis (Davis, 2009; Galaczi, 2008). In addition, researchers have established a relationship between joint collaboration and individual speakers' use of specific linguistic features (e.g., first- and second-person pronouns, *wh*-questions, *that* deletion) which are considered to be markers of personal involvement (McDonough & Uludag, 2021). In the current study, the raters assigned consistent ratings for collaboration, characterized by equality and mutuality, across individual and shared performance. Such consistency might have occurred because during the warm-up, the speakers were prompted to exchange their opinions and engage in a discussion by relating personal contributions to their interlocutor. The speakers' personal involvement and collaborative behaviors, such as confirming comprehension and responding to an interlocutor's collaborative attempts, were salient to the raters when evaluating both individual and shared performances. In that sense, our findings respond to the concerns about assessing collaboration as an individual achievement (Taylor & Wigglesworth, 2009).

Importantly, our rubric did not include a dimension for nonverbal ability, although the raters used both audio and video recordings for assessing oral interactions. Comparison of audio and video conditions did not reveal significant difference in these raters' assessments of paired interactions in the absence of rubric criteria and explicit rater training on the nonverbal aspect of communication. (Appendix B). However, previous rater perception studies reported a positive relationship between paired speaking test performance and speakers' use of body language, eye contact, facial expressions, head nods, and gestures (Ducasse & Brown, 2009; Ducasse, 2013; May, 2011). Given that research is limited in this area with methodological variation (Nakatsuhara et al., 2021), future studies need to look into the impact of rating modality in terms of the salience of visual input for the raters.

Implications

Because the difference between shared and individual performance in this study was evident for discourse management only, this leaves researchers with a number of questions about how interactional features of discourse are evaluated in paired speaking tests. To maximize the usefulness of test results and broaden the representation of the domain of interactional competence, it is key to identify how individual speakers contribute to organization of discourse during pair interactions. For language program administrators, a main take-away is this: When implementing paired speaking tasks in classroom-based assessment situations, it is essential to determine in which contexts it could be beneficial to evaluate students' discourse management collectively as opposed to individually. For example, pairings of low and high proficiency test takers might expose the gap between individual speakers' contributions to discourse. In addition, task types, such as negotiation, decision making, or information exchange, might impact the extent to which discourse is co-constructed between test takers. Drawing on the subcategories for the discourse management domain included in our rubric, course instructors could train L2 speakers to use various discourse features to develop their interactional competence across different task types.

Although we found that the difference between shared versus individual assessment of paired oral task performance is manifested in the rubric category of discourse management only, several considerations might limit the extent to which these findings might apply to other similar assessment contexts. First, although all conversation partners had equal social status (as university students), the target pairs were selected based on their self-reported standardized test scores, without consideration of their L2 proficiency. To clarify how the speaking partners' varying linguistic abilities might impact their task performance, future research needs to incorporate a larger sample of participants and investigate language proficiency more systematically. Second, it is likely that the speakers did not experience test anxiety because they carried out paired interactions in a research lab which was specifically designed for data collection. However, the presence of video cameras and audio recorders might have negatively affected the authenticity of their conversations and encouraged ritualized talk (e.g., *okay*, *I know*, *yeah*) (He & Dai, 2006; Luk, 2010). Besides, since paired oral interactions were not administered for assessment purposes in this study, it is unclear whether the results would be similar for classroom-based or high-stakes assessment situations where students might experience test anxiety. Thus, future research is needed to corroborate empirical investigations of paired speaking tests in various contexts to validate their usefulness as authentic assessment instruments.

In addition, despite establishing acceptable levels of reliability for the rubric, we note that its validity needs to be confirmed using qualitative analysis of rater behavior and

performing psychometric analysis (Lynch & McNamara, 1998; Kim, 2015). To provide practical implications for assessment and teaching, it is critical to define the specific constructs that fall within interactional competence. This requires validating rubric criteria and descriptors which define performance at different levels of interactional proficiency.

---

*1 There were no significant intercorrelations among distinct rating criteria, which suggest that the raters were not subject to the halo effect.

Funding statement

Funding for this study was provided by a grant provided to the first author by the Canada Research Chairs program (Grant number 950-231218).

Acknowledgments

We would like to thank the research assistants for their assistance with data handling, coding, and rating: Dalia Elsayed, Roza van Lieshout, Rachael Lindberg, Libing Lu, and Jie Qiu.

References

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (ed.), *The Companion to Language Assessment* (pp. 1301–1322). https://doi.org/10.1002/9781118411360.wbcla070

Beltrán, J. (2016). The effects of visual input on scoring a speaking achievement test. *Studies in Applied Linguistics and TESOL, 16*(2), 1–23.

Berry, V. (2007). *Personality differences and oral test performance*. Peter Lang.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. L. Erlbaum

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

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

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

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*, 367-396. https://doi.org/10.1177/0265532209104667

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26,* 423–443. https://doi.org/10.1177/0265532209104669

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

Együd, G., & Glover, P. (2001). Oral testing in pairs – a secondary school perspective. *ELT Journal, 55*(1), 70–76. https://doi.org/10.1093/elt/55.1.70

ffrench, A. (1999). *Study of qualitative differences between CPE individual and paired test format (Internal UCLES EFL report).* University of Cambridge Local Examinations Syndicate.

Galaczi, E. D. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English*. [Unpublished doctoral dissertation]. Teachers College, Columbia University, USA.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*, 89–119. https://doi.org/10.1080/15434300801934702

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219-236. https://doi.org/10.1080/15434303.2018.1453816

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*, 370–401. https://doi.org/10.1191/0265532206lt333oa

Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-65.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal, 87*(1), 90-107. https://doi.org/10.1111/1540-4781.00180

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239-261. https://doi.org/10.1080/15434303.2015.1049353

Kormos, J. (1999). Simulating conversations in oral proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*, 163–188. https://doi.org/10.1177/026553229901600203

Kumpulainen, K., & Mutanen, M. (1999). The situated dynamics of peer group interaction: An introduction to an analytic framework. *Learning and Instruction*, *9*, 449–473. https://doi.org/10.1016/S0959-4752(98)00038-3

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65,* 127-159. https://doi.org/10.1111/lang.12115

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, *7*, 25-53. https://doi.org/10.1080/15434300903473997

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180. https://doi.org/10.1191/026553298674579408

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge University Press.

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly*, *5*, 313-335. https://doi.org/10.1080/15434300802457513

Limesurvey GmbH. / LimeSurvey: An Open Source survey tool /LimeSurvey GmbH, Hamburg, Germany. URL http://www.limesurvey.org

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, *5*(1), 95–110.

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

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

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

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS Speaking Test performances. *Language Assessment Quarterly. 18*(2), 83-106. https://doi.org/10.1080/15434303.2020.1799222

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills : A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal, 24*(1), 15–31. https://doi.org/10.1177/003368829302400102

McDonough, K., & Trofimovich, P. (2019). *Corpus of English as a Lingua Franca Interaction* (CELFI). Concordia University.

McDonough, K., & Uludag, P. (2021). Individual and shared assessment of ESL students' paired oral test performance: Examining rater judgments and lexico-grammatical features. In W. Crawford (Ed.), *Multiple perspectives on learner interaction: The corpus of collaborative oral tasks* (pp. 69-91). Mouton De Gruyter.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446–466. https://doi.org/10.1093/applin/18.4.446

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal, 59*, 287 - 297. https://doi.org/10.1093/elt/cci057

Norton, J. (2013). Performing identities in speaking tests: Co-construction revisited. *Language Assessment Quarterly, 10*(3), 309-330. https://doi.org/10.1080/15434303.2013.769549

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*, 277–295. https://doi.org/10.1191/0265532202lt205oa

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143-154. https://doi.org/10.1016/S0346-251X(02)00002-7

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*, 119–58. https://doi.org/10.1111/1467-9922.00179

Styles, P. (1993). *Inter-and intra rater reliability of assessments of 'live' versus audio-and video- recorded interviews in the IELTS Speaking test*. UCLES Research report.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275–302. https://doi.org/10.1177/026553220101800302

Taylor, L. (2001). The paired speaking test format: Recent studies. *Cambridge ESOL Research Notes, 6*, 15–17.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*, 325-339. https://doi.org/10.1177/0265532209104665

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*, 489-508. https://doi.org/10.2307/3586922

Vo, S. T. (2019). *Effects of task types on interactional competence in oral communication assessment* (Doctoral dissertation). Iowa State University. ProQuest Dissertations and Theses Global

Appendix A

Individual and shared rubrics

Individual Rubric					
Performance domains	Subcriteria	Score Levels			
		4 Always	3 Mostly	2 Sometimes	1 Rarely
Discourse management	There is a clear organization of ideas; the speech is consistent and the ideas flow in a logical order.				
	The speaker uses full range of discourse markers (linkers, transition words) to organize and connect ideas.				
Collaborative communication	The speaker initiates, responds, and engages with the other speaker's ideas (asks questions, asks for opinions etc.)				
	The speaker develops the topic making equal contributions to the task.				
Content development	The speaker delivers content that is informative, creative, and relevant to the task.				
	The speaker provides reasoning for their arguments (through examples, explanations etc.).				
Language	The speaker produces accurate and complex structures with no major grammatical problems.				
	The speaker uses a wide range of topic related vocabulary accurately (form) and appropriately (meaning).				
	The speaker produces language with no hesitation and self repetition.				

Shared Rubric					
Performance domains	Subcriteria	Score Levels			
		4	3	2	1

		Always	Mostly	Sometimes	Rarely
Discourse management	There is a clear organization of ideas; the speech is consistent and the ideas flow in a logical order.				
	The speakers use full range of discourse markers (linkers, transition words) to organize and connect ideas.				
Collaborative communication	The speakers initiate, respond, and engage with each other's ideas (asking questions, asking for opinions etc.)				
	The speakers develop the topic making equal contributions to the task.				
Content development	The speakers deliver content that is informative, creative, and relevant to the task.				
	The speakers provide reasoning for their arguments (through examples, explanations etc.).				
Language	The speakers produce accurate and complex structures with no major grammatical problems.				
	The speakers use a wide range of topic related vocabulary accurately (form) and appropriately (meaning).				
	The speakers produce language with no hesitation and self repetition.				

Appendix B

Rubric scores by rating stimuli

	Audio	Video
Category	*M* (*SD*)	*M* (*SD*)
Discourse management	5.7 (.88)	5.6 (.99)
Collaborative communication	5.7 (.95)	5.9 (.73)
Content development	6.0 (.88)	5.9 (.87)
Language accuracy and complexity	8.2 (1.2)	8.0 (1.2)

One-way ANOVAs were carried out to compare the rubric scores awarded to audio and video recordings. The results yielded no statistically significant differences between group means for any of the rubric categories: discourse management, $F(1, 58) = .169$, $p =$

.68, collaborative communication, $F(1, 58) = 1.03$, $p = .31$, content development, $F(1, 58) = .410$, $p = .52$, and language accuracy and complexity, $F(1, 58) = .577$, $p = .45$.