Apples – Journal of Applied Language Studies
Vol. 13, 1, 2019, 151–174

Assessing interactional skills in a paired speaking test: Raters' interpretation of the construct

Linda Borger, University of Gothenburg

The operationalization of interactional competence (IC) within the paired speaking test format allows for a range of interactional skills to be tested. However, in terms of assessment, challenges are posed with regard to the co-constructed nature of IC, making investigations into raters' perceptions of the construct essential to inform test score interpretation. This qualitative study explores features of IC that raters attended to as they evaluated performances in a paired speaking test, part of a Swedish national test of English as a Foreign Language (EFL). Two groups of raters, 17 EFL teachers from Sweden, using national standards based on the Common European Framework of Reference for Languages (CEFR), and 14 raters from Finland and Spain, using CEFR scales, rated six audio-recorded paired performances, and provided written comments to explain their scores and account for salient features. The findings of the content analysis indicate that raters attended to three main interactional resources: topic development moves, turn-taking management, and interactive listening strategies. As part of the decision-making process, raters also considered the impact of test-takers' interactional roles and how candidates' performances were interrelated. In the paper, interaction strategies that were perceived as more or less successful by raters are highlighted. The findings have implications for our understanding of raters' operationalization of IC in the context of paired speaking tests, and for the development of rating scales and guidelines that reflect the social dimensions of the construct.

Keywords: construct operationalization, interactional competence, Swedish national test of English, paired speaking test, Common European Framework of Reference for Languages (CEFR)

1 Background

Following the communicative approach to language teaching and learning, paired and group orals[1] have become widespread in second/foreign (L2)[2] language speaking assessments. With this development, the construct underlying speaking tests has expanded to include social aspects of language use (McNamara & Roever, 2006). The concept of interactional competence (IC) was first introduced by Kramsch (1986) and has been developed in slightly different versions in several subsequent publications (Hall, 1993, 1995; A. W. He & Young, 1998; Jacoby & Ochs, 1995; Young, 2000, 2008, 2011). At the heart of the conceptualization of IC lies the notion that communication is co-constructed and context-dependent and

Corresponding author's email: linda.borger@gu.se
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä

http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201903011694

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

therefore varies with the interactional practice (A. W. He & Young, 1998; Young, 2000). These two characteristics hold obvious challenges for assessment, making investigations into raters' perceptions of IC essential to inform test score interpretation. This paper first reviews theoretical and empirical work of IC in relation to paired/group speaking assessments. It then explores features of IC that raters attended to as they evaluated performances in a paired speaking test, part of a Swedish national test of English as a Foreign Language (EFL).

McNamara (1997) defines two main perspectives from which a speaking construct for L2 assessment can be conceptualized: a *psychological*, focusing on mental activities within the individual speaker, and a *social-behavioural*, recognizing the role of the social context and its effect on interaction. As McNamara (1997) and others (e.g., Chalhoub-Deville, 2003; Johnson, 2001; Young, 2000) have pointed out, conventional approaches to L2 assessment based on Hymes' (1972) theory of *communicative competence*, two of the most influential being Canale and Swain (1980) and Bachman and Palmer (1996), represent a primarily psychological-cognitive conceptualization of interaction, making them unsuitable as a framework of IC. Young (2011) notes that the fundamental difference between communicative competence and IC is that "an individual's knowledge and employment of these resources is contingent on what other participants do" (p. 430). In line with this, Galaczi and Taylor (2018) advocate a socio-cognitive perspective (Weir, 2005) of IC, according to which spoken interaction is seen:

> both as a cognitive and a social interactional trait, with emphasis not just on the knowledge and processing dimension of language use, as seen in the Bachman and Palmer (1996) model, but also on the social, interactional nature of speaking, which has as its primary focus the individual in interaction. (p. 221)

This definition was taken as a starting point for the present study for understanding how IC is interpreted by raters and represented in assessment scales.

Empirically, it has been found that the construct underlying paired/group speaking tests can tap into a fuller range of speech functions than the examiner-led interview format (Brooks, 2009; ffrench, 1999; Kormos, 1999; O'Sullivan, Weir, & Saville, 2002). Peer interaction tasks also provide the potential for a more balanced power relationship between the participants (Iwashita, 1996; Kormos, 1999; Lazaraton, 2002). However, given the complex interaction patterns displayed in peer interaction, the task type has attracted criticism. There is a range of studies that have investigated interlocutor variables that may influence test-taker performance, including, for example: *gender* and *cultural background* (O'Loughlin, 2002; O'Sullivan, 2002), *personality* (Berry, 2007; Nakatsuhara, 2009; Ockey, 2009), *proficiency levels* of the interlocutors (Davis, 2009), and *acquaintanceship* among interlocutors (Chambers, Galaczi, & Gilbert, 2012). Overall, however, the results are mixed indicating that a linear relationship between personal characteristics and scores cannot be claimed.

Another issue involves the joint construction of performance, which makes test-takers' performances inextricably linked. This poses questions of fairness and reliability as a test-taker's scores may differ if they perform a similar task but with a different interlocutor (Davis, 2009; May, 2009). A key question under debate is whether the variability inherent in social interaction should be considered construct-irrelevant variance (Messick, 1989) or part of the speaking construct.

From a socio-cognitive (Weir, 2005) or interactionalist perspective (Chapelle, 1998), the latter position is advocated.

2 Literature review

Lam (2018) noted that there is a growing body of literature on IC in the two fields of L2 learning and L2 assessment. However, the way IC is conceptualized differs in the two research strands. Whereas the L2 assessment literature focuses on interactional conduct and management, thereby treating formal aspects of language as a separate area, the L2 learning literature investigates context-sensitive conduct and how linguistic resources are used to accomplish interactional actions (see, e.g., Pekarek Doehler & Berger, 2018). Keeping this important distinction in mind, as well as the implications it has for construct conceptualization, this literature review will, for reasons of space, only give reference to works dealing specifically with L2 assessment.

The operationalization of IC in paired/group speaking tests has been investigated from two main perspectives in the L2 assessment literature; (1) through discourse-based studies of test-taker interaction and (2) through rater orientation studies. Discourse-based studies using conversation analytic (CA) methodology (e.g., Galaczi, 2008, 2014; Gan, 2010; Gan, Davison, & Hamp-Lyons, 2009; Lam, 2018; Nakatsuhara, 2013) have contributed to a more fine-tuned description of IC. Some common interactional features investigated are:

- *topic negotiation* – how test-takers initiate, develop and shift topics
- *interactional contingency* – how test-takers respond to and engage with each other's ideas
- *turn-taking management* – for example, how transitions between turns are carried out
- *listener support strategies* – how test-takers show listener involvement and support

A feature that has received less attention is *task management.* Using peer interaction data from national tests of English at the lower secondary school level in the Swedish context, Sandlund and Sundqvist (2011, 2013) investigated how test-takers managed task-related trouble, and how this affected topic management and subsequent ratings. Furthermore, the progression of IC skills in testing contexts has been explored by Galaczi (2014) and Gan (2010). Results demonstrate that high-scoring test-takers engage contingently with each other's ideas, manage the conversational floor cooperatively, and provide frequent listener support. Lower scoring students displayed less mutual and cooperative behaviors, indicating difficulty in keeping both the speaker and listener role active simultaneously.

Additionally, Galaczi (2008) distinguished three main interactional patterns in the discourse produced by 30 pairs in the First Certificate in English (FCE) speaking test, based on dimensions of 'mutuality' and 'equality' (Damon & Phelps, 1989; Storch, 2002): (1) *collaborative*, with test-takers working cooperatively (high mutuality) and contributing to talk equally (high equality), (2) *parallel*, with speakers taking on 'solo' roles; both initiating topics but not engaging to any great extent with each other's ideas, and (3) *asymmetric*, characterized by a dominant and a passive speaker. In terms of scoring, collaborative groups achieved the highest scores for the criterion *Interactive Communication*, whereas parallel groups achieved

the lowest and asymmetric groups falling in-between. Galaczi concluded that asymmetric interaction posed the greatest challenges for assessment.

The issue of asymmetric interaction and the implications for the separability of scores in paired assessment was further explored in May (2009), who used both rater data and excerpts of test-taker discourse to highlight an example of a candidate who was given a lower score when involved in asymmetric interaction with a partner of a similar proficiency level, and a higher score when interacting with a partner of a higher proficiency level, which resulted in a collaborative pattern. May (2009) argued that awarding shared scores for IC may be "one way of acknowledging the inherently co-constructed nature of interaction in a paired speaking test" (p. 397). However, this issue is debated and Nakatsuhara (2013), for example, maintained that joint scores would be unfair in cases of asymmetric interaction where one test-taker tries hard to invite and involve more quiet partners but fails to do so. In addition, collaborative interaction may be difficult to reconcile with the focus on individual performance in high-stakes assessments (L. He & Dai, 2006; Luk, 2010), which highlights the importance of task design and framing in order to maximize the elicitation of IC features.

Raters' interpretation of IC has been explored in studies using verbal protocol analysis. Ducasse and Brown (2009) collected verbal report data from 12 teacher-raters who viewed 17 paired performances in the context of an Australian university Spanish beginner-level course. Three interaction parameters were identified: (1) *non-verbal interpersonal communication*, suggesting that body language and gaze are part of the IC construct, (2) *interactive listening*, including candidates' manner of displaying comprehension and giving supportive feedback, and (3) *interactional management skills*, including management of topics and turns.

Complementary results have been demonstrated in rater studies by Orr (2002) in the context of the Cambridge English speaking tests and May (2009, 2011), who reported on features of IC that were salient to four trained raters of 12 paired speaking test in an English for Academic Purposes (EAP) university context. Both Orr (2002) and May (2011) found that raters heeded many aspects that were not explicitly described in the rating criteria, such as references to non-verbal behavior, and comments involving comparisons of candidates' performances. May (2009, 2011) also discovered that raters, although being trained to assess test-takers' performances individually, viewed key features of the interaction as mutual achievements (e.g., the authenticity and quality of the interaction).

To sum up, the literature review shows that different research methodologies have been used to explore IC in the context of paired and group oral assessments, leading to a comprehensive description of the construct. Nevertheless, despite consensus about the co-constructed nature of IC, "there is [...] much ambivalence and debate among language testing scholars on how to deal with it" (Lam, 2018, p. 380). Investigations into raters' interpretation of IC may help inform the discussion about the validity issues associated with the construct, as well as contribute to the development of rating scales, rater guidelines and training materials that reflect its multi-componential nature. Considering the local and situational nature of IC, investigations in different assessment contexts is desirable. The present study is set within a school context, which differentiates it from previous rater studies, which have mainly been performed in the context of tertiary level education and large-scale international tests.

Through a content analysis of raters' written justifications of scores, the present paper aims to highlight features of IC that raters attend to while awarding scores in a paired speaking test. The two research questions addressed are as follows:

1. What features of interactional competence do raters attend to as they judge performances in a paired speaking test?
2. According to the raters, what characterizes more or less successful interaction?

Additionally, a tentative analysis of the nature of, and relationship between, the descriptors for IC in the Swedish EFL performance standards and CEFR scales was made.

3 Setting, participants, and data

3.1 Assessment context

The study reported here is part of a larger research project (Borger, 2018) which explored different aspects of validity evidence in relation to a paired speaking test, part of a national test of English at the upper secondary level of the Swedish school system. In Sweden, national tests are not final exams but have an advisory function in teachers' decision-making regarding students' final grades. However, as the results should "be taken into special consideration" (Swedish Ministry of Education and Research, 2017, pp. 22–23), the tests are regarded as distinctly high-stakes. The national tests are centrally developed but internally rated by teachers, who are provided with assessment guidelines and benchmarked examples, in addition to national standards. The typical national test of English comprises three subtests: a speaking test, a writing test, and a section focusing on reception; i.e. listening and reading comprehension.

Furthermore, the present study is set within the context of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, 2018). The CEFR emphasizes the importance of interactional ability in its division of speaking into two skills: production and interaction, and the inclusion of, for example, a general scale for "qualitative aspects of spoken language use" (Council of Europe, 2001, pp. 28–29) and three specific scales for interaction strategies (ibid., pp. 86–87). The document has had a significant impact on the development of the national syllabuses for foreign languages (Swedish National Agency for Education, 2011), which are related to, and have been textually aligned to, the common reference levels in the CEFR. For example, the minimal pass level of the English 6 course, used as an example in the present study, is intended to correspond to a low B2.

3.2 Raters

Raters from two contexts related to the CEFR participated in the study: (1) formally qualified teachers of English in Sweden (n = 17) from eleven different upper-secondary schools in two national regions, and (2) CEFR raters from Finland (n = 7) and Spain (n = 7). The methodological choice to include external raters was motivated by the opportunity this provided to make a small-scale, tentative comparison between the national EFL standards in the Swedish school context and the reference levels in the CEFR (see Borger, 2014).

Due to the design of the study, regarding for example time required for raters to participate, a combination of convenience and purposive sampling was used to select the participants (see Borger, 2018). The Swedish teachers had varying teaching experience, on average 12 years. They were all well acquainted with and had experience of rating the national tests of English. There were four males and

thirteen females in the Swedish group. The raters from Finland and Spain were all EFL education professionals (working at schools/universities and/or ministries) with a high level of familiarity with the CEFR, as well as previous experience using CEFR scales. There were two males and twelve females in this group.

3.3 Candidates and speaking task

Six audio-recorded paired conversations were used in the study, collected from a pre-testing round of the national test for upper secondary course English 6 (entrance level intended to correspond to CEFR level B2.1)[3]. The test-takers were around 17 years old and each pair consisted of a male and a female student. The conversations were chosen to be representative of different proficiency levels. The speaking task involves a discussion between two students about a given topic. In the first part, focusing on both production and interaction, test-takers are required to present a short text they have read during the 15-minute preparation time and discuss it with their partner. In the second part, focusing on interaction, test-takers discuss different aspects of the topic[4]. In the guidelines to the test, teachers are explicitly instructed to keep in the background and let the students control the conversation. The time allowed for the task is about 15 minutes.

3.4 Rating scales

Test-takers' performances were scored holistically. The Swedish teacher raters and the CEFR raters used separate rating scales, both, however, based on the CEFR. The teachers from Sweden used the national EFL performance standards (Appendix 1) and generic analytic assessment factors (Appendix 2), intended to provide additional support for teachers in making their holistic judgement. This study is specifically concerned with the operationalization of IC. As can be seen in Appendix 1 and 2, IC is described in quite general terms, such as as students' ability to *express themselves clearly with fluency, and with adaptation to purpose, recipient and situation* and to *use communicative strategies to develop and advance the conversation and to solve linguistic problems, e.g., through reformulations, explanations and clarifications*. It is worth mentioning that the National Agency for Education provides additional support materials, in which the EFL standards are explained and clarified in relation to the CEFR (Swedish National Agency for Education, 2011). It may therefore be expected that the Swedish raters were familiar with more explicit descriptions and examples of interaction strategies.

The CEFR raters used scales from *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment – A Manual* (Council of Europe, 2009, pp. 184–186), consisting of a global scale (p. 184) and an analytic scale (pp. 185–186)[5]. The descriptors for "Interaction" draw attention to concepts such as '*can ask and answer questions*', '*keep conversation going of his/her own accord*', '*initiate, maintain and close*' conversation, '*help the discussion along*', '*confirming comprehension*', '*taking turns*', '*inviting others in*', '*keeping the floor*', and '*relate his/her own contributions skillfully to those of other speakers*'.

3.5 Data collection

To examine and explore raters' cognitive processes in language testing, think aloud protocols are increasingly used (Ericsson & Simon, 1993; Green, 1998; Suto, 2012). In this study, due to constraints on time and the amount of data possible to

encode, the raters were asked to verbalize their rating decision in writing and comment on specific features that contributed to their judgment. The data were collected during one-day seminars. After an introduction, comprising information about the study and some basic training, the raters independently listened to the six conversations and provided written comments. The raters from the Swedish context were given the choice of writing either in English or Swedish. As would be expected, raters differed in quantity and type of comments made, which is typically found in other rater report studies too (Ducasse & Brown, 2009; May, 2011). The comments had a mean value of 72 words per performance, range 9–250 words.

4 Method of analysis

The data were explored using content analysis, which can be used both qualitatively and quantitatively (Galaczi, 2013; Hsieh & Shannon, 2005). The present study mainly employed qualitative content analysis. However, frequency counts for the coding categories were computed to serve as an index of the salience of these features (see Krippendorff, 2013). The analysis was carried out deductively on the basis of existing theory and prior research, a so-called directed approach (Hsieh & Shannon, 2005).

To validate the analyses, and to reduce coder subjectivity, two researchers with PhDs in applied linguistics agreed to function as co-coders. They received explicit instructions and some basic training. The first cycle of data analysis involved identifying sections of the holistic rater comments that concerned IC. This initial segmentation was carried out independently by the main researcher and the two co-coders on 45% of the total dataset, resulting in an inter-coder agreement between the main researcher and the co-coders of 76% and 89% respectively. Cases of disagreement, particularly pertaining to the coding of raters' comparisons between test-takers' performances, and the development of test-takers' own ideas as part of IC, were resolved through discussions.

Next, the segments were carefully read through by the main researcher to identify specific categories of features of IC that raters perceived as successful or less successful. On the basis of this, a draft set of coding categories was devised, based on the categories used in Ducasse and Brown (2009) and Galaczi (2008, 2014) (see Literature Review above). In addition, the features of performance described in the rating scales used in the study, including the three descriptor scales for interaction strategies in the CEFR (Council of Europe, 2001, pp. 86–87), proved useful for the more detailed description of sub-categories. In the next step, the segments were further divided into units of analysis, i.e. features focusing on the same theme or idea (Green, 1998), and coded using the draft coding protocol. To illustrate how comments were segmented and coded, an excerpt is provided below. The boundaries between units of analysis are delineated with a '/' and followed by an abbreviation of the coding category:

> /She listens to what the male is saying and as the conversation develops she acknowledges his thoughts. /IL:C
> /and even adds her own opinion to the subject at hand. /TDM:TC
> /She even puts the question back to him for further discussion. /TDM:TQ
> [Rater 1, Sweden]

The first unit of analysis was coded under the main category "IL", which refers to *interactive listening strategies*; and was further coded under the subcategory *confirmations*, indicated by "C". The second unit was coded as "TDM", referring to *topic development moves*; and was further coded under the subcategory "TC", which refers to *topic cohesion.* The final comment involves a reference to *topic development moves* (TDM); and was further coded under the subcategory *topic questions* (TQ).

Forty-five per cent of the total dataset was double coded by the main researcher and the two co-coders. A satisfactory level of inter-rater agreement between the coding of the main researcher and the two co-coders was achieved, 85% and 86% respectively, at the main category level. Cases of disagreement were resolved through discussions, and a final set of coding categories, and an agreed set of units of analysis and codings, were thus reached. The final coding scheme included five main themes, listed below (see Appendix 3 for full coding scheme).

- Topic development moves
- Turn-taking management
- Interactive listening strategies
- Interactional roles
- Additional comments on interaction

In the last step, the researcher independently segmented and coded the remaining dataset (55%) according to the final coding scheme. The software NVivo 11 was used to organize and analyze the data.

5 Results and discussion

Each of the five categories emerging form the content analysis is defined and discussed below and examples from the rater comments are given by way of illustration. The presentation highlights what raters described as successful and less successful interaction strategies. After this, the frequency counts undertaken for each of the five main categories will be presented and a brief comparison between the Swedish and CEFR rater groups in terms of the salience of IC features will be made.

5.1 Topic development moves

The first category, *topic development moves*, was drawn from comments on test-takers' efforts to stimulate and develop the content of the conversation as an interlocutor; in other words, strategies used by test-takers to help the development of topics/ideas. The category was further divided into two subcategories.

The subcategory *topic cohesion* described how test-takers initiated, developed and connected topics, and to what degree this was done in a collaborative manner that helped the discussion along. Raters noted that test-takers who developed their own ideas, justified their opinions and exemplified moved the conversation forward in a constructive way, as in the example: '*She makes several good observations and uses examples to develop her thoughts, which moves the topics along*' [Rater 10, Sweden]. Similarly, extending a partner's ideas was viewed positively, as illustrated in: '*Really good at adding perspectives to the male student's topics*' [Rater 12, Sweden], and '*She provides opportunities for her partner to continue with the topic*

in a really natural way' [Rater 7, CEFR]. Furthermore, raters found the ability to introduce new topics and connect topics with what had previously been said to be successful strategies:

> His partner's contributions being so short and few and far between, he manages to keep the conversation going by introducing new ideas and examples. [Rater 5, CEFR]
>
> Refers back to previous discussions regularly and with fluency. [Rater 8, Sweden]

Raters commented more negatively when test-takers failed to engage with their partner's ideas and contributions, or did not develop their own ideas, as exemplified in: '*Her contribution to the conversation is limited and she doesn't hook onto many comments made by her partner. She delivers her own opinions but doesn't add much motivation'* [Rater 1, Sweden].

The second subcategory, *topic questions*, comprised comments on test-takers' use of questions. Galaczi (2014) remarked that questions perform two main functions in spoken interaction: they manage the distribution of turn-taking and help extend topics under development. The latter was illustrated in comments where raters noted that test-takers used questions as a collaborative strategy to move the discussion forward, as in:

> she contributes to interaction with questions [Rater 9, CEFR]
>
> She poses good and relevant questions to her partner. This moves the conversation forward and contributes to interesting discussions. [Rater 11, Sweden, tr.][6]

However, raters were critical of the overuse of questions, when test-takers asked questions without contributing to topic development:

> Tends to ask partner a lot of questions instead of contributing more with his own viewpoints. [Rater 14, Sweden]
>
> She does not contribute much to the conversation. She keeps asking "What do you think?" as she struggles to find something to say. [Rater 5, CEFR].

Similar to what was found in May (2011), overuse of questions could be interpreted by raters as "an attempt by the candidate to deflect attention from own weaknesses" (p. 138).

The importance of topic management as an essential component of IC is consistent with the findings in both rater studies (Ducasse & Brown, 2009; May, 2011; Orr, 2002) and CA analyses (Galaczi, 2014; Gan et al., 2009; Lam, 2018; May, 2009; Nakatsuhara, 2013). The results of the current study can also be related to the progression of IC skills shown in Galaczi (2014), where it was demonstrated that test-takers at the lower CEFR levels mainly developed their own topics and rarely contributed to the development of their partner's ideas. At higher proficiency levels, in comparison, the ability to develop topics in a mutual manner and across several turns was shown to increase.

5.2 Turn-taking management

The second category emerging from the rater comments was *turn-taking management*. The majority of comments pertained to *turn-taking strategies*. In

particular, raters commented on test-takers' ability to initiate and maintain discourse. It was perceived positively when test-takers took the initiative, were active and kept the discussion going, as in the following example:

> She interacts with ease and skill with natural turntaking, referencing, … She is engaged in keeping the conversation going on. [Rater 9, CEFR]

Conversely, when test-takers showed difficulties in moving the discussion along, by taking limited initiative, this was regarded as negative: '*He also does little to keep the discussion going but mostly just waits for his partner to respond to his comments*' [Rater 10, Sweden]. Furthermore, raters found the ability to intervene appropriately important for successful interaction. When a test-taker dominated turn-taking by interrupting their partner to gain the floor, not allowing enough space to respond, this was viewed in a negative light, as exemplified in: '*She does, however, interrupt her partner a lot and does not allow him to say much*' [Rater 17, Sweden].

The subcategories *turn tempo/length* involved comments on how natural and smooth transitions between turns were perceived to be, thus highlighting 'confluence' (McCarthy, 2010) or 'interactional oral fluency' (Sato, 2014). Raters used expressions such as: '*conversational fluency*', '*natural turn-taking*', '*maintain the conversation flowing*' and '*discussion flows nicely and smoothly*'. In addition, raters noticed aspects connected to turn length and speed of response:

> There is well timed turn taking in the conversation which gives balance in the exchange of information. [Rater 1, Sweden]

> He struggles to get a word in and when he does, he is too slow for his partner so she jumps in. [Rater 5, CEFR]

The fact that raters noticed 'interactional fluency' as an indication of successful interaction is corroborated in Galaczi (2014), where a progression in turn-taking management was shown: "the ability to start a turn after a latch/overlap was found to increase with proficiency level and as learners became more efficient at simultaneously decoding their partner's utterance, composing their contributions and projecting the end of the turn" (Galaczi, 2014, p. 572). Successful test-taker interaction thus resembles casual, real-life conversation where speakers often start turns in an overlapping manner. This has implications for the matching of students, as it is indicated that both candidates in a pair need to have reached a certain developmental stage to be able to manage smooth transitions.

5.3 Interactive listening strategies

The third interaction feature identified in the rater comments was *interactive listening strategies*, which included comments on test-takers' efforts to display attention or engagement while listening. In this category, receptive and productive skills overlap, as test-takers switch between the role of listener and speaker. Listening as part of a test-taker's interactive skills was divided into three subcategories. The first subcategory, *confirmations*, was the largest and comprised comments on test-takers' ability to actively monitor partner's speech and confirm mutual understanding. Raters commented positively on candidates' ability to give

feedback and comment on their partner's contributions. This was seen to help the discussion along, as illustrated below:

> They listen to each other and give very natural feedback signals and display respect for each other as communication partners. [Rater 14, CEFR]

> He listens to what she says and moves on from there. Helps develop the conversation. [Rater 4, Sweden, tr.]

Raters noted that responses to a partner's statements should be substantive, as in: '*she also adds constructive comments and valid points when responding to her partner*' [Rater 10, Sweden]. Conversely, it was perceived more negatively if test-takers responded minimally, or provided formulaic expressions, for example just agreeing or disagreeing with partner's contributions without justifying further: '*Almost no discussion whatsoever, as he just tends to say "I agree" to all the points made*' [Rater 8, Sweden]. It was also pointed out that supportive listening using minimal responses could be a strategy to mask lack of comprehension, which was discussed by Ducasse and Brown (2009) in their rater study too. The following example illustrates this:

> To start with I thought she was listening actively and was interested in what he was saying, but I noticed after a while that she repeated everything he said. [Rater 11, Sweden, tr.]

In addition, another type of listener support noted by raters was *back-channeling*, i.e. brief verbal or non-verbal signals of engagement provided while the other speaker maintains the floor. This was perceived as positive for the development of the conversation when used and negative when lacking, as demonstrated in the examples below:

> Keeps up the conversation by using "conversation particles" such as "oh, yes, yeah, uhum, uhm, ok". [Rater 13, CEFR]

> He does not make any confirmatory/acknowledging sounds or comments which results in the conversation never really getting started. [Rater 6, Sweden, tr.]

As has been pointed out in other rater and CA studies (Ducasse & Brown, 2009; Galaczi, 2014; Lam, 2018), the difference between more substantive responses and backchannelling is that the latter does not provide evidence of comprehension. Ducasse and Brown (2009) therefore speculated that:

> …raters might potentially jump either way with such behaviours, interpreting them positively (providing interactional support) or negatively (a lack of comprehension). Back-channels could also be seen (positively) as an aspect of strategic competence, whereby a listener encourages the speaker to continue until they reach a point where they understand sufficiently to re-join the conversation. (p. 438)

It is also interesting to note that Galaczi (2014) found that at the lower and intermediate CEFR levels, test-takers provided limited listener support, whereas at C1 and C2, they had developed their ability to act as supportive listeners by using substantive confirmations of comprehension. As was the case for turn-taking skills, the development of interactive listening skills seems to be related to

test-takers' increased efficiency in simultaneously monitoring their partner's speech and constructing a response.

The second subcategory, *clarifications*, was drawn from comments on test-takers' efforts to respond to interactional trouble[7] by asking for or giving clarification or help, as in the example: '*She asks her partner to clarify a point in his text and reacts appropriately to his answer*' [Rater 5, CEFR]. Offering or requesting clarification was viewed as a functional strategy that helped move the conversation forward: '*Also when she doesn't understand her partner's description, she asks him to explain it for her, which is a good dialogue technique*' [Rater 8, Sweden]. Raters also noticed candidates offering help by providing, or filling in, a word or expression the partner had trouble with.

Another kind of listener engagement was shown in the last subcategory *flexibility*, which pertained mainly to the Swedish raters' comments on test-takers' ability to accommodate speech to the situation and recipient; an aspect emphasized in the national performance standards. In addition, this category included references to test-takers' efforts to paraphrase or reformulate ideas to achieve mutual understanding, perceived as a positive strategy: '*Reformulates ideas and makes sure that the other understands and supports them too*' [Rater 4, CEFR].

5.4 Interactional roles

The fourth category, *interactional roles*, included comments referring predominantly to one of the six conversations, in which raters perceived the interaction pattern to be 'asymmetric' (Galaczi, 2008). One of the test-takers, in this case the female student, was dominating the interaction, whereas the other, male student, had a more passive role. The female student in this pair received the highest average score of the twelve test-takers in the sample; she was clearly a proficient speaker, as was evident in the rater comments. However, raters also noticed that she dominated the conversation by interrupting the male speaker, not giving him enough time to take the floor. A dominant interactional role was related to both turn-taking and listening skills, as illustrated in the following rater comments:

> Takes over too much and sometimes interrupts partner and perhaps she doesn't listen as much as she should. [Rater 3, Sweden]

> She has a hard time interweaving her contribution into the joint discourse with fully natural turntaking. She doesn't give him a chance. Is this good or bad? She has the language, but perhaps not the personality..? C1 or C2? For the benefit of the student, C2, with reservations. What sort of interaction do we want to encourage? [Rater 3, CEFR]

Some raters noted that the dominant interactional style of the female speaker interfered with her partner's capacity to demonstrate his full potential, and could thus be perceived as disadvantageous. However, opinions differed. As illustrated below, the interactional style of the dominant, more proficient speaker, could also be seen as having a positive effect on her partner's performance. This is similar to what was found in May's (2009) study in which raters did not agree on whether to penalize or compensate test-takers for their roles in asymmetric interaction. The issue of separability of scores is thus particularly relevant in cases of asymmetric interaction. The following examples illustrate raters' somewhat different views of asymmetric interaction in the same pair:

> During the rest – he is repeatedly interrupted by the female student, who speaks too much. It is hard to hear his full range, since he does not "fight" her verbally, he lets her takes over. [Rater 12, Sweden]

> I felt that this speaker was somewhat disadvantaged due to a domineering partner. I would have liked to have heard more. [Rater 14, Sweden]

> I have a feeling that if the boy had been interacting with a less proficient counterpart, his production might have been closer to B2, but as he had such a fluent and at times dominant partner, he was able to stretch himself beyond what he would have been able to do if he had been conversing with a less proficient interlocutor. [Rater 3, CEFR]

In Galaczi's (2008) study, a collaborative interaction style was associated with the highest interactive communication scores. It is therefore interesting to note that the dominant speaker in the current study received the highest average score in the sample, and her more passive partner also performed above average. The complexity of asymmetric interactions was further highlighted in Davis (2009), where one of three candidates who participated in both collaborative and asymmetric interactions received a higher score when engaged in dominant asymmetric interaction. Davis (2009, p. 387) concluded that, "although collaborative interaction was generally associated with higher level examinees and scores, there did not appear to be a penalty in terms of score when an examinee's interlocutor was unable to maintain a collaborative interaction". Clearly, further research is needed to explore this complexity, which will be returned to in the conclusion.

5.5 Additional comments on interaction

Finally, the last category, *additional comments on interaction*, included three subcategories. The first subcategory contained general comments with no specific interaction feature pointed out, comprising both positive and negative examples: '*Interacts well enough, contributes to a good discussion*' [Rater 9, Sweden] and '*shows restricted interactive ability*' [Rater 10, CEFR].

Raters also made comments on their *rating decision*. This was done predominantly by the CEFR raters, who referred to the CEFR descriptors, for example: '*Interacts with ease in the joint discourse (C2)*' [Rater 1, CEFR]. In addition, raters commented on how test-takers performed in *comparison*, or in relation to one another, which has also been demonstrated in previous rater orientation studies of paired speaking (May, 2011; Orr, 2002). Raters noted similarities and differences, and compared and contrasted test-takers' performances, as illustrated below:

> Nice interaction – nuanced discussion, but doesn't come through as much as his partner. [Rater 9, Sweden]

> Both of them jump from one topic to the other and make a few comments but there is no a real discussion. [Rater 7, CEFR]

In some cases, raters referred to the pair as one entity, for example: '*they seem to have a good team work*' [Rater 2, Sweden], thus acknowledging the co-constructed nature of interaction. Raters also commented on how well-matched test-takers were in terms of proficiency level: '*This pair seems to be at comparable levels of proficiency*' [Rater 3, CEFR]. Finally, there were comments in this subcategory on

the interdependence of test-takers' performances, once again highlighting the impact test-takers may have on each other's performances in a paired speaking test. There were both positive and negative examples, as illustrated below:

> I feel he could have performed better with a more collaborative partner with better contributions. [Rater 5, CEFR]

> Sometimes overwhelmed by his partner but it also helped him to maintain a high level. [Rater 8, Sweden]

The tendency of raters to make comparisons can be considered a form of relative judgement; that test-takers' performances are being assessed in comparison with one another instead of in relation to the rating criteria. Both May (2011) and Orr (2002) voiced concerns about this based on their rater studies. However, comparing test-takers' performances can also be a seen as part of raters' operationalization of the IC construct. Young (2011) stresses that IC is jointly constructed by the participants. In light of this, considering test-takers' performances in relation to the other group participants seems justified.

5.6 Quantitative results of coding

In Table 1 below, the frequency counts and percentage of coding categories for the Swedish and CEFR rater groups and the total are provided. As can be seen, there are differences with regard to the salience of IC features between the teacher raters from Sweden, who were guided by the national performance standards, and the group of external CEFR raters, who were guided by CEFR scales[8], suggesting that the rating scales emphasize slightly different facets of IC. Whereas *topic development moves* and *interactive listening strategies* were more frequently mentioned by the teacher raters from Sweden, the CEFR raters made proportionally more comments in the categories *turn-taking management* and *additional comments on interaction*. However, it is worth emphasizing that this is mainly a qualitative study, describing general patterns of raters' perceptions of IC. The quantification of results therefore needs to be interpreted with caution, not least due to the purposeful sampling and limited sample size.

Table 1. Frequency counts and percentage of coding categories for interaction across raters

Main coding categories	**Swedish raters (*n* = 17)**	**CEFR raters (*n* = 14)**	**Sw + CEFR (*N* = 31)**
Topic development moves	197 (39%)	58 (19%)	255 (31%)
Turn-taking management	43 (9%)	61 (20%)	104 (13%)
Interactive listening strategies	137 (27%)	56 (18%)	193 (24%)
Interactional roles	36 (7%)	31 (10%)	67 (8%)
Additional comments on interaction	90 (18%)	103 (33%)	193 (24%)
Total	503	309	812

The limited scope of this paper does not make it possible to discuss the quantitative results in detail, but a few remarks will be made. To start with, *topic development moves* were commented on to a larger extent by the raters from

Sweden, which may be explained by the wording in the rating scales. In the Swedish assessment materials (Appendices 1 and 2), the development of ideas is highlighted in terms of complexity and variation – that test-takers should be able to *give different examples and perspectives of the topics discussed* and *use communicative strategies to develop and advance the conversation*. Perhaps this is less explicit in the CEFR descriptors for interaction (Council of Europe, 2009, pp. 184–186), expressed as, for example, the ability to *interact on a wide range of topics*, *express points of view*, and *relate one's own contribution to those of other speakers*. Furthermore, the reason why the Swedish raters commented more frequently on *interactive listening strategies* may be attributed to the fact that the national performance standards emphasize the ability to *accommodate speech to purpose, recipient and situation*, and rater comments pertaining to this feature were incorporated under this category.

Conversely, the CEFR raters made proportionally more comments on *turn-taking management*, which may be attributed to the emphasis given to this facet in the descriptors, for example *initiate, maintain and close discourse*, *preface remarks in order to get or to keep the floor*, *take his/her turn when appropriate* and *interweave contribution into the joint discourse*. In the Swedish rating scale, turn-taking is not explicitly mentioned, but is incorporated as an element of *functional* or *communicative strategies*. It is thus indicated that the less explicit reference to aspects of turn-taking management in the Swedish descriptors may have had an impact on raters' attention to this feature of IC.

The analysis also revealed that CEFR raters made proportionally more comments in the category *additional comments on interaction*, the difference being most noticeable for the two subcategories *rating decision* and *comparisons*. An explanation for this may be that the CEFR raters were rating a test that was 'new' to them, and that this led to more references to descriptors and comparisons between test-takers' performances. The Swedish raters, in comparison, rated a test that is a recurring part of their professional practice. It is also a distinctly high-stakes test which has consequences for students' final grades, which may lead to a stronger individual focus.

6 Conclusion and implications

The present study explored features of interaction that raters attended to while judging performances in a paired speaking test. The content analysis of rater comments identified three main interactional resources employed by test-takers: *topic development moves*, *turn-taking management*, and *interactive listening strategies*. These were seen as contributing to successful interaction when used in a collaborative and mutual manner, with test-takers actively monitoring and responding to their partner's speech. In comparison, less successful interaction strategies were characterized by weaker alignment between test-takers and a lower degree of collaborative and interpersonal moves. As part of the decision-making process, raters also considered the impact of test-takers' *interactional roles*, and how candidates performed in relation to one another. In other words, the evidence collected from raters' comments substantiates the benefits of the paired speaking test format in terms of eliciting a rich repertoire of test-takers' interactional skills, while also highlighting the challenges for validity.

The findings of the study correspond, in a broad sense, to what has been shown in other studies of paired oral testing and further emphasize the need to take

contextual as well as individual factors into account, thus including the variability inherent in social interaction as part of the construct. Galaczi and Taylor (2018) argue that there is a "need for a speaking test construct defined in social terms that taps into interaction and goes beyond a purely psychological conceptualisation, and that encompasses a certain degree of unpredictability and less control than the individual interview format speaking test" (p. 225). The key question, then, is how variability in paired/group speaking tests can be captured without compromising reliability. Here, further research is obviously needed, regarding, for example, IC task design (see, e.g., van Batenburg, Oostdam, van Gelderen, & de Jong, 2018) and effects of technological advancements such as automated scoring systems (see, e.g., Bernstein, Van Moere, & Cheng, 2010) and computer-mediated paired/group speaking tests (see, e.g., Nakatsuhara, Inoue, Berry, & Galaczi, 2017).

Additionally, the findings of the current study suggest that the complexity involved in assessing asymmetric test-taker interaction is an area that warrants further investigation. The test instructions in the Swedish school context state that the teacher should "point out to students that they should help each other to keep the conversation going" and "encourage students to give each other roughly equal speaking opportunity" (Swedish National Agency for Education, 2017, p. 17). It is further emphasized that it is important that students are given the opportunity to display their ability. However, although guidelines exist, it seems clear that more explicit information and directions are needed concerning the potential effects of personal interlocutor variables and how to address this issue when assigning scores, including advice on the matching of students (the latter already present in the national tests for lower secondary school).

The present study holds implications for large scale testing as well as for teaching and continuous assessment of oral proficiency performed by teachers in classrooms. The topic investigated, namely how raters construe L2 interaction and interpret standards and criteria, has obvious connections to other research areas, including *language teacher cognition* (Borg, 2006), the growing field of *language teacher assessment literacy* (Fulcher, 2012; Hill, 2017; Tsagari et al., 2018), and *learning-oriented language assessment* (Jones & Saville, 2016). With reference to the latter two areas, Hamps-Lyons (2017) argues that "there should be assessment literacy materials available for learners/test-takers as well as for teachers and test raters/interlocutors" (p. 101). Data from rater and discourse-based studies may provide a useful foundation for such activities and support materials. As pointed out by Galaczi (2014), "a more precise delineation of IC is important for L2 teachers and learners, as it provides guidance for the development of interactional skills in a communicative classroom" (p. 555).

Similar to what has been demonstrated in previous rater orientation studies (May, 2011; Orr, 2002), this study shows that raters' interpretation of IC provides a more comprehensive view of the construct than was reflected in the rating scales. As Brooks (2009) noted, "there was a lot more going on in the paired format than the rating scale captured" (p. 361). In light of this, rating scales have to be further developed, representing the reciprocal and mutually constructed characteristics of interaction, as well as illustrating the progression of IC skills more clearly. In addition, the issue of individual versus joint scores for IC needs further exploration; however, it seems clear that interlocutor effects have to be considered when results are interpreted, and guidelines for raters as well as for teachers have

to be elaborated, including conceptually grounded reasoning as well as commented examples.

Finally, the main limitations of the current study need to be acknowledged. First, as this is mainly a qualitative study, convenience and purposive sampling was used with a total of 31 raters. Obviously, this has implications for generalizability. However, since the findings are well in line with previous research they are of clear interest in further analyses of issues concerning L2 interaction. Second, as the intention was to keep the investigation as close to the authentic rating procedures as possible, only audio-recorded paired speaking tests were used. Previous research (Ducasse & Brown, 2009; May, 2011; Nakatsuhara, 2011) has given strong indications that non-verbal features, such as body language, facial expressions and gaze, are part of the IC construct, and this was not possible to investigate in the present analyses. This is an avenue for further research in the context of the paired speaking task used in the Swedish national test of English.

In spite of these limitations, the study has hopefully contributed some useful insight to the assessment, teaching, and learning of interactional skills, both in high-stakes and classroom contexts.

Endnotes

[1] Paired and group oral test formats are speaking tests in which test-takers interact with each other, not with an examiner.
[2] In this paper, the term L2 is used to refer to both second and foreign languages.
[3] Tasks and items for the national EFL tests are extensively piloted and pretested. For more information on test construction and development, see Erickson and Åberg-Bengtsson (2012).
[4] On the National Assessment Project webpage, sample tests are provided for reference: https://nafs.gu.se/prov_engelska/engelska_gymn/exempel.
[5] Table C2: Oral assessment criteria grid on p. 185 (Council of Europe, 2009) corresponds to CEFR Table 3: Common Reference Levels: Qualitative aspects of spoken language use (Council of Europe, 2001, pp. 28–29).
[6] The Swedish raters were given the choice of writing either in English or Swedish. Examples from rater comments translated from Swedish are henceforth marked 'tr.'
[7] The concept used in CA is repair (Schegloff, 2000).
[8] Although an important aspect of rating, intra-group differences are not explored in the present study. The focus is on raters' general perceptions of the construct of IC, and not on individual differences.

References

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

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

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

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. London: Continuum.

Borger, L. (2014). *Looking beyond scores. A study of rater orientations and ratings of speaking*. Licentiate thesis, University of Gothenburg, Sweden. Retrieved from http://hdl.handle.net/2077/38158

Borger, L. (2018). *Investigating and validating spoken interactional competence: Rater perspectives on a Swedish national test of English.* Doctoral dissertation, University of Gothenburg, Sweden. Retrieved from http://hdl.handle.net/2077/57946

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

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

Chambers, L., Galaczi, E. D., & Gilbert, S. (2012). Test-taker familiarity and speaking test performance: Does it make a difference? *Research Notes, 49*, 33–40.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York: Cambridge University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating examinations to the Common European Framework: a manual*. Retrieved from http://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assesment. Companion volume with new descriptors*. Retrieved from http://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 13*(1), 9–19.

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

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

Erickson, G., & Åberg-Bengtsson, L. (2012). A collaborative approach to national test development. In D. Tsagari & I. Csépes (Eds.), *Collaboration in Language Testing and Assessment* (pp. 93–108). Frankfurt am Main: Peter Lang.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (revised edition)*. Cambridge, MA: MIT Press.

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

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

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

Galaczi, E. D. (2013). Content analysis. In A. J. Kunnan (Ed.), *The Companion to Language Assessment, Vol. 3* (pp. 1325–1339). Chichester: Wiley-Blackwell.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574.

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236.

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

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A Conversation Analytic approach. *Applied Linguistics, 30*(3), 315–334.

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

Hall, J. K. (1993). The role of oral practices in the accomplishment of our everyday lives: The sociocultural dimension of interaction with implications for the learning of another language. *Applied Linguistics, 14*(2), 145–166.

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

Hamps-Lyons, L. (2017). Language assessment literacy for language learning oriented assessment. *Papers in Language Testing and Assessment, 6*(1), 88–111.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Philadelphia: John Benjamins.

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

Hill, K. (2017). Language Teacher Assessment Literacy – scoping the territory. *Papers in Language Testing and Assessment, 6*(1), iv–vii.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolingustics: Selected readings* (pp. 269–293). Harmondsworth: Penguin.

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

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171–183.

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Jones, N., & Saville, N. (2016). *Learning oriented assessment: A systemic approach* (Studies in Language Testing Vol. 45). Cambridge: UCLES & Cambridge University Press.

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*(2), 163–188.

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

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Lam, D. M. K. (2018). What counts as "responding"? Contingency on previous speaker contribution as a feature of interactional competence. *Language Testing, 35*(3), 377–401.

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

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly, 7*(1), 25–53.

May, L. (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*(2), 127–145.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal, 1*(1), 1–15.

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

McNamara, T., & Roever, C. (2006). *Language testing: the social dimension*. Oxford: Blackwell Publishing.

Messick, S. A. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation co-constructed?* Unpublished doctoral dissertation, University of Essex, UK.

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

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt am Main, Germany: Peter Lang.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly, 14*(1), 1–18.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169–192.

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

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33–56.

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*(2), 143–154.

Pekarek Doehler, S., & Berger, E. (2018). L2 interactional competence as increased ability for context-sensitive conduct: A longitudinal study of story-openings. *Applied Linguistics, 39*(4), 555–578.

Sandlund, E., & Sundqvist, P. (2011). Managing task-related trouble in L2 oral proficiency tests: contrasting interaction data and rater assessment. *Novitas – Research on Youth and Language (ROYAL), 5*(1), 91–120.

Sandlund, E., & Sundqvist, P. (2013). Diverging task orientations in L2 oral proficiency tests: a conversation analytic approach to participant understandings of pre-set discussion tasks. *Nordic Journal Of Modern Language Methodology, 2*(1), 1–21.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System, 45*(1), 79–91.

Schegloff, E. (2000). When 'others' initiate repair. *Applied Linguistics, 21*(2), 205.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119–158.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice, 31*(3), 21–30.

Swedish Ministry of Education and Research. (2017). *Nationella prov – rättvisa, likvärdiga, digitala. Prop. 2017/18:14* [National tests – fair, equivalent and digital. Government Proposition, publication number Prop. 2017/18:14]. Stockholm: Swedish Ministry of Education and Research.

Swedish National Agency for Education. (2011). *Kommentarmaterial till ämnet engelska. Om ämnet Engelska* [Subject commentary. About the subject English]. Retrieved from https://www.skolverket.se/download/18.6011fe501629fd150a28916/1536831518394/Kommentarmaterial_gymnasieskolan_engelska.pdf

Swedish National Agency for Education. (2017). *English 6, Spring Term 2017. Lärarinformation – inklusive bedömningsanvisningar till Delprov A Focus: Speaking* [English 6, Spring Term 2017. Teacher information – including scoring guidelines for Subtest A Focus: Speaking]. Stockholm: Swedish National Agency for Education.

Tsagari, D., Vogt, K., Froelich, V., Csépes, I., Fekete, A., Green, A., . . . Kordia, S. (2018). *Handbook of assessment for language teachers*. Retrieved from http://taleproject.eu/. ISBN 978-9925-7399-1-2 (digital).

van Batenburg, E. S. L., Oostdam, R. J., van Gelderen, A. J. S., & de Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing, 35*(1), 75-100.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Young, R. (2000). *Interactional competence: Challenges for validity*. Paper presented at the Annual meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium, Vancouver, BC, Canada.

Young, R. (2008). *Language and interaction. An advanced resource book*. London & New York: Routledge.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426–443). London & New York: Routledge.

Appendices

Appendix 1. Performance standards for oral and written production and interaction, course English 6 (entrance level corresponding to CEFR level B2.1) in Swedish upper secondary school

Grade E	**Grade C**	**Grade A**
*In oral and written communications of various genres, students can express themselves in a way that is **relatively** varied, clear, and **relatively** structured. Students can also express themselves with fluency and **some** adaptation to purpose, recipient and situation. Students work on and make **simple** improvements to their own communications.*	*In oral and written communications of various genres, students can express themselves in ways that are varied, clear and structured. Students can also express themselves with fluency and **some** adaptation to purpose, recipient and situation. Students work on and make **well grounded** improvements to their own communications.*	*In oral and written communications of various genres, students can express themselves in ways that are varied, **balanced**, clear and structured. Students can also express themselves with fluency and adaptation to purpose, recipient and situation. Students work on and make **well grounded and balanced** improvements to their own communications.*
*In oral and written interaction in various, and more formal and complex contexts, students can express themselves clearly with fluency, and with **some** adaptation to purpose, recipient and situation. In addition, students can choose and use **essentially** functional strategies which **to some extent** solve problems and improve their interaction.*	*In oral and written interaction in various, and more formal and complex contexts, students can express themselves clearly, **relative freely** and with fluency, and with adaptation to purpose, recipient and situation. In addition, students can choose and use functional strategies to solve problems and improve their interaction.*	*In oral and written interaction in various, and more formal and complex contexts, students can express themselves clearly, **freely** and with fluency, and with adaptation to purpose, recipient and situation. In addition, students can choose and use **well** functioning strategies to solve problems and improve their interaction, **and take it forward in a constructive way.***

Source: The Swedish National Agency for Education (2011), searchable from https://www.skolverket.se/undervisning/gymnasieskolan/laroplan-program-och-amnen-i-gymnasieskolan/amnesplaner-i-gymnasieskolan-pa-engelska

Appendix 2. Assessment factors provided in Teacher Guidelines for the national test for course English 6 in Swedish upper secondary school

CONTENT

- intelligibility and clarity
- complexity and variation
    - different examples and perspectives
- coherence and cohesion, structure
- adaption to purpose, recipient/interlocutor, situation and genre

LANGUAGE AND ABILITY TO EXPRESS ONESELF

- communicative strategies
    - to develop and advance the conversation
    - to solve linguistic problems, e.g., through reformulations, explanations and clarifications
- fluency and ease of speaking
- range, variation, complexity, clarity and accuracy
    - vocabulary, phraseology and idiomaticity
    - pronunciation and intonation
    - grammatical structures
- adaption to purpose, recipient, situation and genre

Translated from Swedish

Appendix 3. Final coding scheme used to code rater comments pertaining to IC

Final main coding categories	Sub-categories/ Refined categories	Details for second coders: Descriptions/explanations
1. TOPIC DEVELOPMENT MOVES	Topic cohesion	Develop conversation by introducing, extending or connecting topics. Expand and develop 'own/self-initiated' and 'other-initiated' topics/ideas by e.g. giving examples, new arguments, and adding details to the topic Link new topics with previously discussed ones – help focus the talk Keep the development of ideas on course **(Cf. CEFR scale *Co-operating*, 2001, p. 86)**
	Topic questions	The use of questions to move the conversation forward
2. TURN-TAKING MANAGEMENT	Turn-taking strategies	Initiate, maintain and end conversation Being active, taking up space in conversation Intervene in discussion (appropriately) Preface remarks in order to get the floor, or to gain time and keep the floor **(Cf. CEFR scale *Taking the floor (Turntaking)*, 2001, p. 86)**
	Turn tempo/length	How fast/slowly candidates respond Interactional flow; natural/automatic/smooth turn-taking How long/short turns are
3. INTERACTIVE LISTENING STRATEGIES	Confirmations	Commenting on/giving feedback on/following up partner's statements, in order to confirm mutual understanding Agreeing/disagreeing with partner's statement Backchannelling Signals of engagement: Encouraging partner to continue; sound interested; show involvement in partner's contributions **(Cf. CEFR scale *Co-operating*, 2001, p. 86)**
	Clarifications	Giving explanation/clarification Asking for explanation/clarification Offering help Filling a silence/providing missing word **(Cf. CEFR scale *Asking for clarification*, 2001, p. 87)**
	Flexibility	Ability to accommodate speech as a means of expressing it to the situation and the recipient **(cf. CEFR *Flexibility scale*, 2001, p. 124)**
4. INTERACTIONAL ROLES	Asymmetric interaction pattern: Dominant speaker role	Reference to candidate with dominant interactional role; • Dominating conversation, interfering with partner's capacity to demonstrate full potential (usually mentioned negatively) • Managing conversation; supportive role (usually mentioned positively)
	Asymmetric interaction pattern: Passive speaker role	Reference to candidate with passive interactional role
	Parallel interaction pattern: 'Solo' roles	Reference to interaction exhibiting low degree of mutuality and higher degree of equality with test-takers taking on 'solo' roles
5. ADDITIONAL COMMENTS ON INTERACTION	General comments	General comments on interaction, no specific interaction feature pointed out
	Rating decision	Rating decision to do with interaction features/interaction patterns
	Comparisons	How candidates perform in relation or in comparison to one another – similarities, differences Reference to both candidates as an entity

Received December 15, 2017
Revision received October 31, 2018
Accepted February 5, 2019