Assessing Korean ESL Learner’s Interactional Competence through Oral Interviews and Paired Speaking Tasks: A Pilot Study

Hwijung Lee

(Seoul National University)

Lee, Hwijung. 2018. Assessing Korean ESL Learner’s Interactional Competence through Oral Interviews and Paired Speaking Tasks: A Pilot Study. *SNU Working Papers in English Linguistics and Language 16, 101-124.* This paper is a preliminary empirical study. It explores several papers on validating Interactional Competence (IC), a construct that has been largely underrepresented in speaking tests, to understand its theoretical basis. Several complexities regarding IC test design and its evaluation methods are discussed. In this field of research oral interviews and paired speaking tasks are two types of tests widely used. The selection of the tasks is another issue to consider in answering the question namely, whether the scores for a co-constructed interaction should be assessed individually or mutually. A vast number of research focus on individualizing the scores due to issues of practicality. This study devises a scoring criteria that can be used for the mutual assessment of IC per pair in different kinds of tasks. Secondly, this study discusses several reliability issues regarding the rating of IC. Lastly, this study analyzes unique interactional features that distinguish oral interview tasks from paired speaking tasks. **(Seoul National University)**

Keywords: speaking construct, interactional competence, oral interview task, paired speaking task

1. Introduction

Many Korean ESL learners face a completely different challenge even after attaining the highest possible English proficiency score in high-stake tests such as TEPS or TOEFL. In a foreign academic setting where they must participate in group discussions or even small talk, they realize that the communication difficulty experienced is often unrelated to their heavily memorized vocabulary or grammar. Many of the interdiscursive and pragmatic functions of communication are untaught and untested by many accessible high-stake speaking tests, while some exclude the speaking section altogether. Interactional competence (IC) is an underrepresented construct when it comes to designing speaking tests. Communication according to McNamara (1997) is "created by individuals in joint constructions" (Fulcher & Davidson, 2007) but we mistakenly make judgements about interactional skill from test scores that were not intended to measure IC. It is difficult to assume much about interactivity from speech data elicited by computer-based and strictly-timed monologues. What kind of communicative goal do we have in mind when we design speaking tests? This research reviews literatures on IC validation and presents the pilot study that assesses IC through oral interview tasks and paired tasks.

2. Literature Review

2.1 Construct Definition

Interactional Competence (IC) has originally been conceptualized by Kramsch (1986) who questioned the actual speaking construct that question-and-answer format tests meant to assess. She broadened the scope of language speaking proficiency by claiming that the final justification should be to make students "interactionally competent on the international scene".

Celce-Murcia et al.'s (1995) proposed model of communicative competence captures the importance of interpersonal language exchange by further specifying components from Canale's model (1980). In this model *discourse competence*, such as coherence or conversation structure, is at the center. This core competence affects and is affected by *actional competence, sociocultural competence,* and *linguistic competence.* All of these components are encompassed by *strategic competence*. Celce-Murcia's additional component which is *actional competence*, is a mainly "interpersonal exchange" function, which is critical in defining IC. It includes making introductions, identifying oneself, and reacting to interlocutor's speech. This models forms an important theoretical basis for developing assessment criteria for IC. According to Young (2000, 2011) IC is both a "practice-specific behavior" and a "practice-independent, person-specific trait. In other words the test is taken to indicate an underlying trait of the person and the performance is taken to indicate the influence of the context. The complexity in validating IC of an individual lies in identifying all independent and dependent features. May (2011) also claims that at the core of construct definition of IC it is important to consider the co-construction. Understanding an interlocutor's message, responding to the partner, and working cooperatively are all mutual achievement. A shared score would reflect on such aspect but there is the problem of separability of scores especially for high-stakes tests.

2.2 Task Used in Assessing Interactional Competence

Plough et al. (2018) point out that in order to elicit appropriate evidence of IC it is important to take a task-based approach to speaking test design. Some of the most common tasks used in assessing IC are oral proficiency interviews or paired discussions.

2.2.1 Oral interviews

Interviews have widely been preferred because they have been characterized as natural, relaxed conversations. However, it is still important to remember they are still a form of language assessment. Even though we attempt to collect close-to-"natural conversation" language samples from interviews, as Lazaraton (1992) claims, the "interaction initiation responsibilities come from outside of the interaction itself". Moreover, she adds that interviews could not be entirely natural for they are still tests under time constraints (Lazaraton, 2002). Indeed, the "rule of the games" for speaking interviews place the candidate in a position where they should use all possible opportunities in order to produce as much language sample as possible rather than responding truthfully (Norton, 2013). Another problem posited by several authors in interview co-construction is the interlocutor variation and corresponding bias (Brown, 2003; Nakatsuhara, 2008; May, 2009). For instance, Norton's (2013) discourse analysis on oral interview tests demonstrate that even with a scripted interlocutor frame *interdiscursivity* could cause an interlocutor to adapt a more "teacherly, supportive" style. When discussing the rigidness of interlocutor frame there will always be a conflict between issues of generalization and practicality with those of authenticity. Inevitably, individual variation would affect not only the candidates in talk but also raters assessing the interaction as well. Roever et al.'s (2018) research on language proficiency interviews (LPI) shows that often other-initiated repair by a candidate may not always be a problem of lexical knowledge but rather one of unexpected topic shift or question format. Mechanism for changing topics, asking questions, and the resulting sequential organization could vary according to interviewer style.

2.2.2 Paired Tasks

Other studies have focused on assessing IC through paired tasks. The greatest advantage these possess is the more interaction-sensitive locality and more collaborative and even authentic speech samples (Galaczi, 2014). However, paired tasks also possess a loophole when it comes to rating. Ducasse et al. (2009) collected verbal reports of raters to analyze their qualitative judgement. Raters preferred paired tasks for their "assesability" of IC due to the equal flow of conversation. Yet, their study points out that some paralinguistic features, such as gaze or body language, as well as listening ("supportive listening") were features that were assessed. Moreover, it seemed that rater's already have a notion of what successful interaction should look like. In May's study (2011) raters viewed the ability to ask the partner for their opinion as a positive feature of IC. Yet, they were critical of candidates who did not express their own opinion as much as they asked their partners. While oral interviews possess problems of interlocutor bias paired tasks have problems of rater bias.

2.2.3 IC assessment methods

Several studies validate IC using different rating methods. Youn (2015) uses a data-driven scoring rubric created from analyzing salient features from actual paired conversations. This criteria therefore has an intensive focus on conversational skills such as *sensitivity to situation, engaging with interaction,* and *turn organization.* In another study, Batenburg et al. (2018) use both analytic rating and holistic rating to assess IC. The results from analytic scoring reveal that some interactional abilities such as *meaning negotiation* and *correcting misinterpretation* stand out in rater's judgement of IC. Also, they suggest that linguistic accuracy and interactional ability are stand-alone criteria and the rubric may therefore fail to explain a candidate's *self-supporting* or *other-supporting* style in interaction. These studies guide future studies' different scoring methods of paired tasks, however, they focus on individualizing the scores due to practical issues. Though seemingly less applicable it is still crucial to consider assigning shared scores. One of the greatest difficulty faced by raters is that there are some clearly inseparable features and mutual achievements (May, 2009). Many ongoing discussions still center on determining the ideal method to capture IC.

2.3 Interactional Competence Scoring Criteria: The CAP Scale

Wang (2015) developed an interactional competence rating scale (the CAP scale) based on the presence or absence of three types of patterns: collaborative, asymmetric, and parallel. It places the collaborative pattern on the highest bands (5-6) when candidates "contribute equally to the conversation and interact cooperatively". What distinguishes a 5 from a 6 is the level of task completion. Asymmetric pattern (3-4) is next on the scale for interactions in which one candidate assumes a more dominant role while the other is more passive. The lowest scores are given to parallel pattern (1-2) when candidates "have equal access to the conversational floor and the development of the interaction, but do not work cooperatively". However, the author points out two concerns regarding the scoring criteria. The CAP scale demonstrated low correlation among raters and also this rubric gave task completion a higher weigh than collaboration at certain band levels. Due to this limitation, it is essential to consider whether task completion is a criteria that should be assessed in an IC rubric. Furthermore, it is questionable whether asymmetry truly stands opposite to collaboration in spoken interaction. To hypothesize, a high degree of asymmetry may be perfectly acceptable depending on the task type. For instance, in the role-play task used in Youn's (2015) research the candidate must fulfill a request from the interlocutor, who plays the role of a professor. It is it likely for such task to have an imbalance in turn-taking and for the uneven power hierarchy to be reflected in conversation. This expected asymmetry would not necessarily translate to a lack of collaboration. As May (2009) argues, most interactions that take place in university settings are asymmetrical.

3. Research Questions

This study will focus on an area less explored: assessing a shared score per interaction. The vast literature suggests there is a need to capture the mutuality and co-construction of conversations although the specifics on how to carry this out remain unanswered. Additionally, this study will use both oral interviews and paired tasks because it is essential to consider the locality of the target language use domain and how the task type could affect the interaction assessed. The questions that guide this

preliminary research are the following:
(1) What kind of scoring criteria should be devised to mutually assess IC per pair?
(2) Which features contribute to rater variability in such assessment of IC?
(3) What are some unique features captured by candidates paired tasks compared to oral interview tasks?

4. Methods

4.1. Candidates

A total of 13 Seoul National University undergraduate and graduate students, seven female and six male, participated in this study. All candidates had less than six months of experience in an English speaking country. Their English proficiency scores (taken within two years) were converted referring to the official conversion table from the TEPS website (https://www.teps.or.kr/InfoBoard/ConversionTable#). Their scores ranged from TEPS 450 to 990. Five candidates (two female and three male) took part in an oral interview task and eight candidates (five female and three male) in paired discussion tasks.

4.2. Interlocutor

There were two untrained interlocutors, one male and one female, who took part in the oral interview tasks. Both interlocutors were English L2 speakers with a high proficiency level (TEPS score above 900) with a background in applied linguistics. In this preliminary study a training session has been omitted due to time constraints. However, the interlocutors were given a set of instructions about the purpose of the study and a set possible interview topics.

4.3. Materials

The tasks consisted of oral interviews or paired speaking interaction: (1) one oral interview task (free discussion topic) (2) paired tasks (free discussion topic and role-play) Each of the interactions, interview or paired, took a minimum of five minutes to maximum ten minutes. All interactions were audio-recorded by the researcher. For paired tasks, a pair completed both Task A and Task B consecutively. Candidates orally received general instructions from the interlocutor (for oral interviews) or the researcher (in the case of paired-tasks) along with an instructions sheet in Korean. At the end of all interactions the candidates took a survey that asked their perceived level of task difficulty for future reference. In the case of paired tasks candidates were to fill out a question that asked their perceived level of their partner's English proficiency level.

4.4. Scoring Rubric

4.4.1. General Proficiency

Two scoring rubrics were used. A General Proficiency (GP) rubric (Appendix) was used to holistically rate all candidate's general proficiency scores. The highest band was a five and the lowest a one (a score of zero meant no data available for assessment). There were five criteria: grammar, vocabulary, pronunciation/fluency, contents delivery, and turn organization. Some of the components were modified from previous research (Youn, 2015; Park, 2017) and others were added by the researcher. The criteria reflect the components of linguistic, discourse, and sociocultural competence from Celce-Murcia's communicative competence model (1995).

4.4.2. Interactional Competence

A holistic scoring rubric, which is a modified version of Wang's (2015) CAP scale was used to assess IC (Appendix). The two components to be individually assessed in this rubric were Collaborative Pattern (CP) and Asymmetric Pattern (AP). The highest score possible for each CP and AP was a four, the lowest was a 1. A score of 0 meant no interaction was present for assessment.

4.5. Raters

There were a total of three raters in this study including the researcher (two female and one male). One of the raters was an English native speaker and the other two including the researcher were L2 Each raters listened to the audio recordings which they could refer to multiple times if necessary. The raters evaluated the General Proficiency for all 13 candidates. A CP and AP score was given to each interaction: one per each oral interview, and one per each of the two tasks in the case of paired interactions. In other words, the CP and AP scores were mutual evaluations for each pair.

4.6. Conversation Analysis

Conversation analysis (CA) transcription conventions (Have, 2004) were used in order to transcribe audio files for qualitative analysis of the interactions. Only data from free discussion task were transcribed to compare patterns in oral interviews and paired discussions.

5. Results

5.1. Ratings

The data in Table 1 and Table 2 below are the mean scores of all three raters for the oral interview and paired tasks.

Table 1. Oral interview task CP, AP, GP raters mean scores

Candidate	Collaborative Pattern (CP)	Asymmetric Pattern (AP)	General Proficiency (GP)
1	1 (SD = 0)	3.33 (SD = 0.58)	2 (SD = 0)
2	3.33 (SD = 0.58)	2.33 (SD = 1.15)	4 (SD = 1)
3	2 (SD = 1)	3 (SD = 0)	3.33 (SD = 0.58)
4	3.33 (SD =1.15)	2.33 (SD = 1.15)	4.67 (SD = 0.58)
5	3.33 (SD = 0.58)	2.33 (SD = 0.58)	3.33 (SD = 0.58)

Table 2. Paired-task raters mean scores

Pair	Candidate	Task A		Task B		GP
		CP	AP	CP	AP	
1	6	4 (SD= 0)	1 (SD = 0)	3.33 (SD= 0.58)	2.33 (SD= 0.58)	4.33 (SD= 0.58)
	7					3.33 (SD= 0.58)
2	8	2.33 (SD= 0.58)	2.67 (SD= 0.58)	1 (SD = 0)	4 (SD = 0)	4.33 (SD= 0.58)
	9					2.33 (SD= 0.58)
3	10	1.33 (SD= 0.58)	3.33 (SD= 0.58)	2.33 (SD= 0.58)	2.67 (SD= 0.58)	2.33 (SD= 1.53)
	11					3 (SD= 0)
4	12	2.33 (SD= 0.58)	3 (SD = 1)	3 (SD = 1)	2.33 (SD= 0.58)	2.67 (SD= 0.58)

	13					5 (SD= 0)

Additionally, Kripendorff's alpha correlation coefficient was calculated using SPSS. The data of all three raters for Collaborative Pattern ratings was a 0.630 and for General Proficiency a 0.636. Scores of 0.700 or higher is considered significant. Although the scores for CP and GP are not at the significant level, given the data size, it is uncertain whether this could yield better results in future research. The rubric is underdeveloped and there was no rater training session as will be reviewed in the discussions section. However, in the case of Asymmetry Pattern the score was a 0.385 which is particularly insignificant. The possible reasons for inconsistent ratings for AP will be discussed in the later sections.

In the case of paired-task ratings, the two different kinds of tasks (free discussion and role-play) prove that IC scores cannot be generalized even to the same pair. Previous studies examined in the literature review discuss the locality and context-specificity of IC. In this data, for instance, pair 2 received a very high AP score (4) for the role-playing task but had a balanced level of scores for CP and AP for the free discussion task.

5.2. Criteria

The correlation coefficient between CP, AP, and GP criteria were compared in order to investigate the linear relationship among them.

Table 3. Correlation between CP, AP and GP ratings for oral interview task

		CP	AP	GP
Collaborative Pattern (CP)	Pearson Correlation	1	-.996*	.855
	Sig. (2-tailed)		.001	.065
	N	5	5	5

Asymmetric	Pearson Correlation	-.996*	1	-.833
Pattern (AP)	Sig. (2-tailed)	.001		.079
	N	5	5	5
General	Pearson Correlation	.855	-.833	1
Proficiency	Sig. (2-tailed)	.065	.079	
	N	5	5	5

*correlation is significant at the 0.05 level (2-tailed)

Table 4. Correlation between CP, AP and GP for paired task A (free discussion)

		CP	AP	GP
Collaborative	Pearson Correlation	1	-.926*	.395
Pattern (CP)	Sig. (2-tailed)		.001	.333
	N	8	8	8
Asymmetric	Pearson Correlation	-.926*	1	-.299
Pattern (AP)	Sig. (2-tailed)	.001		.471
	N	8	8	8
General	Pearson Correlation	.395	-.299	1
Proficiency	Sig. (2-tailed)	.333	.471	
	N	8	8	8

*correlation is significant at the 0.05 level (2-tailed)

Table 5. Correlation between CP, AP and GP for paired task B (role-play)

		CP	AP	GP
Collaborative	Pearson Correlation	1	-.968*	.238
Pattern (CP)	Sig. (2-tailed)		.000	.571
	N	8	8	8
Asymmetric	Pearson Correlation	-.968*	1	-.330
Pattern (AP)	Sig. (2-tailed)	.000		.425
	N	8	8	8
General	Pearson Correlation	.238	-.330	1
Proficiency	Sig. (2-tailed)	.571	.425	
	N	8	8	8

*correlation is significant at the 0.05 level (2-tailed)

The tables above reveal that there is no significant correlation between CP and GP or AP and GP scores. It seems that the GP scores from the rubric in this study did not correlate much with IC. However, there was a significant correlation between the CP and AP criteria. For all three kinds of tasks (oral interview and the two paired tasks) there was a negative linear correlation present. In other words, raters have assessed that in an interaction in which high CP level could be seen there was a low AP level visible and vice versa. Yet, data sample was small and details will be discussed in the next section where CA data is presented.

5.3. Conversation Analysis

5.3.1. Comparison of oral interview and paired free discussion tasks

To compare oral interview and paired formats free discussion tasks were chosen for in-depth conversation analysis. The relatively higher number of turns with relatively low word count could possibly indicate a lot of short turns filled with tokens such as *uh-huh, mm-hm, etc*. The total number of words and total turns were counted from the transcribed files. All five oral interview task recordings ranged from 5:01 to 5:43 minutes length. The paired discussion recordings ranged from 5:06 to 5:41 minutes in length.

Table 6. CA descriptive statistics for oral interview

Candidate	CP	AP	GP	Number of turns	Word count	
					Interlocutor	Candidate
1	1	3.3	2	28	220	157
2	3.33	2.33	4	24	190	248*
3	2	3	3.33	40	281	256
4	3.33	2.33	4.67	32	231	503*
5	3.33	2.33	3.33	32	354	216

*cases in which the candidate's word count exceeds the interlocutor

In Table 6 above it is evident that there were two cases (marked by the asterisk) in which the candidate spoke more words than the interlocutor. Especially, in the case of candidate 4 their word count (503) was nearly double the interlocutor's (231). Compared to candidate 5 who had the same number of turns (32) but a lower word count (216), candidate 4 formed longer monologic responses. The highest general proficiency scores were awarded to both of these verbose candidates. Further study is needed to determine whether a correlation exists between this verbosity and proficiency level or whether the longer and thus assessable speech samples positively biased GP ratings.

Moreover, high CP scores were given to candidates 2, 4 and 5. These candidates had noticeably high level of response tokens and tag-questions. For instance, candidate 4 at the end of their monologic response adds a follow up question to the interlocutor (*Do you like watching mukbangs or do you have any other hobbies?*). The ability to extensively express their opinion and ask the interlocutor's opinion were seen by raters as qualities of high CP.

Overall, AP scores were all over 2.33. More data will be necessary to check whether oral interviews by nature cannot yield an extremely low AP score below 2. For all three interview interactions deemed most collaborative in this data sample, the AP score did not drop below 2.33. Future research should confirm if a truly high CP score could coexist with a relatively moderate or high level of asymmetry (above 2 or 3).

(1) Excerpt 1: Oral interview task (Candidate 1)

(Previous context: explaining Chuseok holiday and yuch nor-i)

46 I: so (.) this yuch that you're telling me is very similar to (0.8) dice?
47 C: °yes° (3.0) °dice° I don't know
48 I: you're not sure? (.) Okay (0.4) and what other kinds of activities do you do (.) in Chuseok?
49 C: mm (9.0) uh (3.0) °like° (4.0) visit (2.0) for grandmother

50 I: mm?
51 C: and (4.0) chat with (3.0) grandmother?
52 I: you chat with your [grandmother?
53 C: [yeh

The excerpt above is from candidate 1 who demonstrated the lowest proficiency level. The interview was rated to have the lowest CP score (1) and a high AP score (3.3) One of the major problems that occurred in the interaction was that it became like an interrogation rather than a conversation. Unlike other candidates this individual had difficulty asking any questions asking the interlocutor's opinion. The interlocutor asks for elaboration about the Korean term *yuch* (line 46) aiding this with a confirmation (*so this yuch… is similar to dice?).* It is revealed later in the talk that the candidate lacks lexical knowledge (line 47: *dice? I don't know*) to be able to any show higher interactive ability. It leads to the interlocutor abandoning the topic completely in line 48, and moving on to another question. To this question the candidate has difficult responding, showing elongated pauses and fragmented speech due to word search (line 49).

Table 7. CA descriptive statistics for paired task

Pair	CP	AP	GP scores A/B	Number of turns	Word count		Question-initiation	
					A	B	A	B
1	4	1	4.33/ 3.33	85	385	337	7	5
2	2.33	3.33	4.33/ 2.33	62	253	241	9	3
3	1.33	3.33	2.33/ 3	42	230	222	3	9
4	2.33	3	2.67/ 5	38	242	403	4	9

Higher level interactive strategies were observed from candidates to whom raters assigned the highest CP score and lowest AP score. For instance, Pair 1 received a CP of 4 and AP of 1. Both candidates received similar levels of GP. Moreover in the survey they showed they had perceived each other to have similar levels of English proficiency.

(2) Excerpt 2: Paired-task A (Pair 1)

(Example of high CP, low AP)

99 A: yea swimming is so good
100 B: yea but nowadays (0.2) the weather is (.) so cold?=
101 A: =<u>yea</u>
102 B: ((laughs))
103 A: ((laughs)) I-I bought the (0.2) th-this ticket? Like I can use swimming pool for a
104 month ticket
105 B: Ohh.
106 A: but (0.2) I (.) don't go
107 B: ((laughs))
108 A: ((laughs)) so just buy and pay but I don't go ((laughs))
109 B: ahh when I uhh when you (.) go to the ticket?

A total of 24 laugh tokens used throughout the interaction which was a high number compared to all other interactions. The laughter in their interaction was not a substitute for filling in gaps or for concealing their linguistic ability. It seems that pairing of similar proficiency level could have played a role in the high level of collaboration. For instance, in line 100 candidate B expresses a comment about the weather to which candidate A responds by acknowledging this with an agreement (line 101: *yea*). A typical trait of their interaction was "supportive listening" in which they demonstrated attention and cooperation. (Ducasse et al., 2008)

5.3.2. Issues regarding proficiency level

The following excerpt is representative of the issues that could be seen in a paired free discussion task where candidates GP score difference was high. The candidates themselves were aware of this discrepancy, as they had replied in the survey: the higher GP candidate perceived the other to have a lower proficiency than them and the lower GP candidate perceived the other speaker to have a higher level of proficiency. This awareness may have affected their interaction as it progressed.

(3) Excerpt 3: Paired-task A (Pair 4)
(Example of a high GP gap)
63 A: yea usually (0.3) uh (.) midfielder
64 B: ah okay ↓ you have to be really good at it (.) so uh (.) isn't it hard? to play soccer
65 when it's really cold like these days↑
66 A: ah yes=
67 B: =how's the weather like today?
68 A: uh (0.4) ºhhh today's cold but uh (0.2) these days cold but today I think like it's not
69 too [much cold?]
70 B: [ahh okay] a::nd (1.2) mm (0.9) do you know any nice place to eat near the campus?
71 A: ah ºhhhh to eat?
72 B: yeah
73 A: uh (.) I usually just eat in campus
74 B: ah okay=

Excerpt (3) above is part of the interaction much later in the free discussion task after candidates have already had time to intuitively estimated each other's proficiency level. Here there the complex process of "natural adaptivity" or "accommodation" (Berwick and Ross, 1992) is displayed by candidate B taking on a more interviewer-like role. In lines 02-03, candidate B shifts the topic from soccer to weather with a preface (*isn't it hard to play soccer when it's really cold like these days*). This closed question format (*isn't it…*) does not leave much room for candidate A to answer but to confirm or disagree with a simple yes or no. Then, candidate B initiates the topic with a question (*how's the weather like today?*) in line 05. The topic has been selected form the prompt but in this case it is evident that A is simply asking B questions to give him the floor. After A responds to the question in lines 06-07, B simply confirms his response (*ahh okay*) rather than giving her own opinions regarding the new topic and moves on to the next question. The sudden shift in topic without B providing her expected portion of the interaction results in B's repair-initiation in line 09 (*to eat?*) which delays his response. Typically, in the paired free discussion tasks of this study both candidates shared each of their opinions for every different subject matters brought to the table. However, in this particular interaction both of the candidates begin with the typically observed pattern but eventually Candidate A takes on an interlocutor-like lead to compensate for the gap in proficiency level in order to keep the interaction flowing. *Compensation* criteria according to Batenburg et al.'s (2018) study, proved to have a low relationship with the IC construct. It may be important to control pairing of candidates to minimize the occurrence of accommodation and compensation because it may affect rater's judgement especially for the AP criteria.

6. Results

In order to answer the first research question, Wang's (2015) CAP scale has been modified for a more detailed assessment of IC. The single IC criteria divided into three band levels (collaborative, asymmetry, and parallel) have been divided into two criteria: collaborative pattern (CP) and asymmetry pattern (AP). Successful collaboration is not necessarily symmetrical, especially in the case of oral interview tasks. Though this study had hypothesized that AP scores for oral interview tasks would generally be higher than for paired tasks, the data size makes such generalization difficult. Moreover, low interrater reliability, particularly for AP, brings issues of rater training and concrete band description to the surface.

A correlation analysis between the CP, AP, and GP constructs revealed that a significant relationship was observed between CP and AP. Relating to the problems of inter-rater reliability for the AP criteria, it is difficult to claim that this correlation is accurate. In the absence of a clear band description for AP and conversation transcriptions it is possible for raters to have relied on the CP construct to make judgements of AP. Furthermore, to better understand the GP features that contribute most to the low correlation between the IC constructs (CP, AP) and GP it may be necessary to conduct analytic scoring. The subscales from analytic rating can provide a detailed interpretation of interactional competence (Van Batenburg et al., 2018).

Lastly, conversation analysis on free discussion tasks of oral interviews and paired talk provide some insight on task differences. The candidate's proficiency level seemed to be an issue for all tasks though there were some differences between oral interviews and paired tasks. In the oral interview task the candidates' level of proficiency seemed to have affected the way they interpreted the test. Those with higher GP levels generally were able to give extended responses and appropriate follow up questions to the interlocutor. These candidates received high CP scores from raters. However, a case of a low proficiency candidate showed that the lower level of linguistic ability may have hindered interactive ability such as asking questions and using acknowledgement tokens (*mmhm, yes*). In paired tasks the highest CP was seen in candidates whom GP levels were similar. However, in pairings where one candidate had a much higher proficiency level than their partners, an interview-like interaction pattern has been observed. Issues regarding the candidate's proficiency level is another hurdle that must be considered for accurate assessment of IC. Clearly, more data is needed to draw a generalizable conclusion.

7. Discussions

7.1. Defining symmetry in IC

This study presented a problem regarding the assessment of AP. There was a particularly low level of interrater reliability. Some of the possible factors for such low agreement rates could be the lack of rater training or vague band descriptions especially for what an asymmetric pattern of conversation should look like. For instance, the band descriptions for AP scores 3-4 was "one candidate assumes a more dominant role and the other a more passive role throughout all or most of the interaction" and scores 1-2 "one candidate assumes a more dominant role and the other a more passive role sometimes". Although the definition of what a "dominant" and a "passive" role have been included this rubric fails to explain the question namely, "dominance" or "passiveness to what degree? It is difficult to operationalize the quality of dominance, but it may be necessary to include some noticeably features from the preliminary conversation analysis. Secondly, the key word that separates band 1-2 from 3-4 (*sometimes, all or most)* refer to the quantity of dominance. Throughout the rating sessions, there were no transcriptions available, but the only the audio files were listened to several times. It may be important for raters to have visual aid that allows for clearer identification of the asymmetry (or symmetry) of the interaction.

7.2. Relationship between general proficiency and interactional competence

The GP scores did not show high correlation with the IC criteria. Although sample size was very small a possible solution for this problem could be to use analytic scoring rubric of GP to further investigate which GP criteria showed (if any) a significant relationship with IC. This leads to the initial question of whether current general proficiency measures predict IC at all.

Another important fact to point out is that the GP scores given by raters already showed a significant gap from candidates TEPS score. For instance, candidate 1 received the lowest GP score from all raters (GP = 2, beginner) but had reported a TOEIC score of 835 (upper intermediate level). This discrepancy between the existing test score and interactional ability poses a problem in collecting participants for the main study. Pairing according to similar proficiency levels is highly recommended (Galaczi, 2014). Yet, existing test scores' low predictability of IC poses a problem in recruiting participants. Perhaps a level test may be necessary to adjust to this problem. An overly low proficiency for oral interview tasks result in an interrogation-like interaction rather than conversation as seen from the CA excerpt. A huge gap of proficiency levels in paired discussion task can result in an interview-like interaction format.

7.3. Other limitations

One rater pointed out that for future ratings laughter should be a controlled feature. She explained that filling in gaps or turns with laughter, though it may leave a good impression for the raters, is a feature to be aware of. Raters should be careful to not award better scores subconsciously due to such features. Arguably, candidates should also be given instructions to raise awareness that substituting words with laughter or other nonlinguistic cues may not necessarily result in higher cores.

References

Brown, Annie. (2003). Interviewer Variation and the Co-construction of Speaking Proficiency. *Language Testing, 20*(1), 1-25.

Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, 1*, 1.

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative Competence: A Pedagogically Motivated Model with Content Specifications. *Issues in Applied Linguistics*, 6(2). Retrieved from https://escholarship.org/uc/item/2928w4zj

Ducasse, A., Brown, A., Taylor, L., & Wigglesworth, G. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment : An advanced resource book / Glenn Fulcher and Fred Davidson.* Abingdon, [England] ; New York: Routledge.

Galaczi, E. (2014). Interactional Competence across Proficiency Levels: How do Learners Manage Interaction in Paired Speaking Tests? *Applied Linguistics, 35*(5), 553-574.

Have, P. (2004). *Doing conversation analysis : A practical guide / Paul ten Have.* (Introducing qualitative methods). London ; Thousand Oaks, Calif.: Sage Publications.

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

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System*, *20*(3), 373-386.

Lazaraton, A., & University of Cambridge. Local Examinations Syndicate. (2002). *A qualitative approach to the validation of oral language tests / Anne Lazaraton*. (Studies in language testing ; 14). Cambridge: Cambridge University Press.

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.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal, 62*(3), 266-275.

Norton, J. (2013). Performing Identities in Speaking Tests: Co-Construction Revisited. *Language Assessment Quarterly, 10*(3), 309-330.

Park, H. (2017). Assessing Interactional Competence of Advanced-Level Korean EFL Students Through L2 Paired Discussion Tasks (Unpublished thesis). Seoul National University, Seoul, South Korea.

Plough, I., Banerjee, J., & Iwashita, N. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing, 35*(3), 325-329.

Roever, C., Kasper, G., Plough, I., Banerjee, J., & Iwashita, N. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331-355.

Ross, S., & Berwick, R. (1992). The Discourse of Accommodation in Oral Proficiency Interviews. *Studies in Second Language Acquisition, 14*(2), 159-76.

Van Batenburg, Eline S. L., Oostdam, Ron J., Van Gelderen, Amos J. S., & De Jong, Nivja H. (2018). Measuring L2 Speakers' Interactional Ability Using Interactive Speech Tasks. *Language Testing, 35*(1), 75-100.

Wang, L. (2015). *Assessing interactional competence in second language paired speaking tasks* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, Arizona.

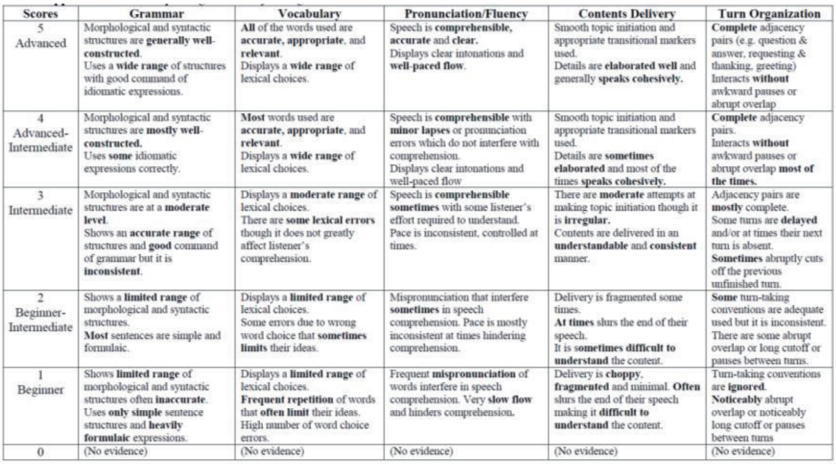
Youn, S. J. (2015). Validity Argument for Assessing L2 Pragmatics in Interaction Using Mixed Methods. *Language Testing, 32*(2), 199-225.

Young, R. F. (2000). Interactional competence: Challenges for validity.

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

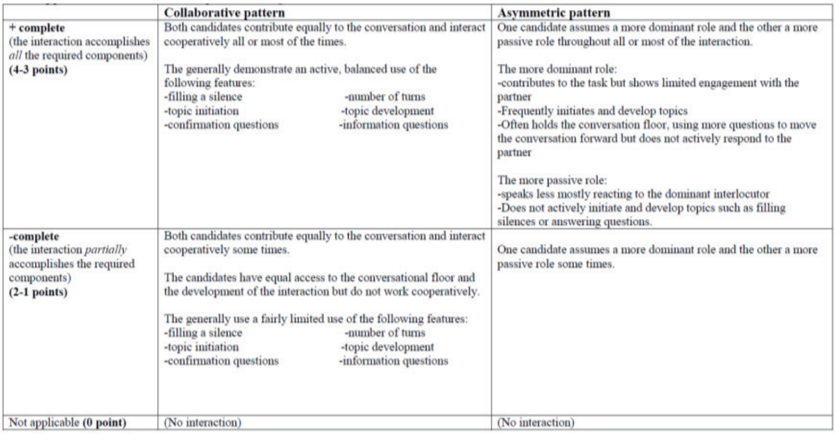
Appendix

General Proficiency Scoring Rubric



Scores	Grammar	Vocabulary	Pronunciation/Fluency	Contents Delivery	Turn Organization
5 Advanced	Morphological and syntactic structures are **generally well-constructed**. Uses a **wide range** of structures with good command of idiomatic expressions.	**All** of the words used are **accurate, appropriate**, and **relevant**. Displays a **wide range** of lexical choices.	Speech is **comprehensible, accurate** and **clear.** Displays clear intonations and **well-paced flow**.	Smooth topic initiation and appropriate transitional markers used. Details are **elaborated well** and generally **speaks cohesively.**	**Complete** adjacency pairs (e.g. question & answer, requesting & thanking, greeting) Interacts **without** awkward pauses or abrupt overlap
4 Advanced-Intermediate	Morphological and syntactic structures are **mostly well-constructed.** Uses **some** idiomatic expressions correctly.	**Most** words used are **accurate, appropriate**, and **relevant**. Displays a **wide range** of lexical choices.	Speech is **comprehensible** with **minor lapses** or pronunciation errors which do not interfere with comprehension. Displays clear intonations and well-paced flow	Smooth topic initiation and appropriate transitional markers used. Details are **sometimes elaborated** and most of the times **speaks cohesively.**	**Complete** adjacency pairs. Interacts **without** awkward pauses or abrupt overlap **most of the times.**
3 Intermediate	Morphological and syntactic structures are at a **moderate level**. Shows an **accurate range** of structures and **good** command of grammar but it is **inconsistent**.	Displays a **moderate range** of lexical choices. There are **some lexical errors** though it does not greatly affect listener's comprehension.	Speech is **comprehensible sometimes** with some listener's effort required to understand. Pace is inconsistent, controlled at times.	There are **moderate** attempts at making topic initiation though it is **irregular.** Contents are delivered in an **understandable** and **consistent** manner.	Adjacency pairs are **mostly** complete. Some turns are **delayed** and/or at times their next turn is absent. **Sometimes** abruptly cuts off the previous unfinished turn.
2 Beginner-Intermediate	Shows a **limited range** of morphological and syntactic structures. **Most** sentences are simple and formulaic.	Displays a **limited range** of lexical choices. Some errors due to wrong word choice that **sometimes limits** their ideas.	Mispronunciation that interfere **sometimes** in speech comprehension. Pace is mostly inconsistent at times hindering comprehension.	Delivery is fragmented some times. **At times** slurs the end of their speech. It is **sometimes difficult to understand** the content.	**Some** turn-taking conventions are adequate used but it is inconsistent. There are some abrupt overlap or long cutoff or pauses between turns.
1 Beginner	Shows **limited range** of morphological and syntactic structures often **inaccurate**. Uses **only simple** sentence structures and **heavily formulaic** expressions.	Displays a **limited range** of lexical choices. **Frequent repetition** of words that **often limit** their ideas. High number of word choice errors.	Frequent **mispronunciation** of words interfere in speech comprehension. Very **slow flow** and hinders comprehension.	Delivery is **choppy**, **fragmented** and minimal. **Often** slurs the end of their speech making it **difficult to understand** the content.	Turn-taking conventions are **ignored**. **Noticeably** abrupt overlap or noticeably long cutoff or pauses between turns
0	(No evidence)	(No evidence)	(No evidence)	(No evidence)	(No evidence)

Interactional Competence Scoring Rubric



	Collaborative pattern	Asymmetric pattern
+ complete (the interaction accomplishes *all* the required components) **(4-3 points)**	Both candidates contribute equally to the conversation and interact cooperatively all or most of the times. The generally demonstrate an active, balanced use of the following features: -filling a silence -number of turns -topic initiation -topic development -confirmation questions -information questions	One candidate assumes a more dominant role and the other a more passive role throughout all or most of the interaction. The more dominant role: -contributes to the task but shows limited engagement with the partner -Frequently initiates and develop topics -Often holds the conversation floor, using more questions to move the conversation forward but does not actively respond to the partner The more passive role: -speaks less mostly reacting to the dominant interlocutor -Does not actively initiate and develop topics such as filling silences or answering questions.
-complete (the interaction *partially* accomplishes the required components) **(2-1 points)**	Both candidates contribute equally to the conversation and interact cooperatively some times. The candidates have equal access to the conversational floor and the development of the interaction but do not work cooperatively. The generally use a fairly limited use of the following features: -filling a silence -number of turns -topic initiation -topic development -confirmation questions -information questions	One candidate assumes a more dominant role and the other a more passive role some times.
Not applicable **(0 point)**	(No interaction)	(No interaction)

Hwijung Lee
Hwijung94@snu.ac.kr