The Journal of English Language Pedagogy and Practice
Vol.11, No.23, pp.180-206, Fall & Winter 2018

Psychometric Characteristics of a Rating Scale for Assessing Interactional Competence in Paired-Speaking Tasks at Micro-level

Milad Ramazani [1], Biook Behnam [2*], Saeideh Ahangari [3]

1,2. Department of English, Tabriz Branch, Islamic Azad University, Tabriz, Iran
*Corresponding author: B.behnam@azaruniv.edu

Received: 2018.2.20
Accepted: 2018.5.27

Online publication: 2019.2.22

Abstract

Developing rating scales for assessing interactional performance is demanding since it is a relatively complicated procedure. The present study investigated the psychometric characteristics of the CAP rating scale (Wang, 2015) for assessing interactional competence at micro-level. To this end, 160 Iranian intermediate EFL learners were selected based on their performance on TOEFL iBT test from a language institute in Tabriz. Four interaction tasks were used to elicit students' performance on interactional competence using the CAP rating scale. Five raters were recruited in the study to assign score to each individual's performance. The participants were pretested and post-tested at the beginning and the end of the term through the same scale. The Pearson correlations were computed in order to estimate the test-retest reliability indices of the scale. In addition, five separate exploratory factor analysis (EFA) through the varimax rotation method were conducted in order to investigate the underlying constructs of the communication functions individually and as a total. The results revealed that the CAP rating scale enjoys a reasonable reliability indices and the four functions i.e. building argument, developing discussion, offering support, and shaping connection can be appropriate predictors of interactional competence. Some pedagogical and assessment implications are presented as well.

Keywords: psychometric characteristics, rating scale, interactional competence, paired-speaking tasks, micro-level

Introduction

A number of communicative competence models have been proposed for the assessment of communicative competence, starting with Lado (1961), Hymes (1972), Canale and Swain (1980) succeeded by Bachman (1990), and more recently Bachman and Palmer (2010). All these models are considered as ability models in nature. By ability model, it is meant that candidate's internal knowledge is the locus of attention and the reciprocities between various characteristics of ability possessed by a specific language user is measured (Kley, 2015). In other words, the locus of attention is on the single user and attempts to elucidate the type of competence a learner requires to apprehend in order to be able to communicate. On the other hand, one of the significant aspect of context which is known as co-construction of discourse which plays an important role in the reciprocal comprehension between interactants has been overlooked (Kley, 2015). However, with the frustration of purely cognitive core of these models along with Bachman's model, new advancements came forth that approach interaction from a more socially oriented prospect (Hall, 1995; Young, 1999). Drawing on the notion that interactional competence and co-construction process are not the ownership of the interlocutors, scholars such as Kramsch, (1986), Young(2008, 2013) together with Jacoby and Ochs (1995) believe that dialogue is inferred as being reciprocally invented between interlocutor in its actual time. Such advancements shaped the emergence of interactional competence.

The degree of co-construction in interactional competence is measured at both the macro-level which is considered as the entire quality of interaction followed by the micro-level which is regarded as interaction features (He & Young, 1998). Specific marks are assigned to each interactant in an oral task, at the micro-level, according to their application of each interaction feature which is categorized as verbal, paralinguistic, and non-verbal (Oskaar, 1990). To provoke interaction performance on either of these levels, there must be oral paired-tasks.

On one hand, with the growing request for the application of communicative language teaching method (CLT) to L2 context, recent decades have noticed increasing attention in paired oral measurement (Taylor & Wigglesworth, 2009; Galaczi, 2013). The rationale behind this increasing consciousness is threefold: an extensive range of interaction abilities are applied by participants and paired speaking tasks offer an instrument for evoking interlocutors' performance in collaboration with their co-participants in conversation (Brooks, 2009; Davis, 2009). Paired

tasks are regarded as an effective plan for educating a variety of linguistic and paralinguistic traits. In the second place, the traditional asymmetry of examiner-examinee power in interviews can be reduced to a great extent (Taylor, 2001). And finally, the centrality of paired actions in communicative contexts such as real tasks situations in the classroom can be addressed by paired speaking tasks (Galaczi, 2013). As a consequence, paired speaking tasks are certified for their validity and prospective positive reaction on L2 language acquisition and are used in this study to elicit verbal interaction features (Ducasse & Brown, 2009).

On the other hand, according to Davies (1990, p. 57), "testing lies at the center of language teaching". It is absolutely impossible to think of any branch of knowledge without considering scaling issues, so the field of testing must be viewed one aspect of measurement for educational assessment. Testing should go hand in hand with teaching.

Noteworthy is that it is demanding to develop speaking rating scales in general and interactional competence scale in particular since it is a relatively complicated procedure due to various grounds and approaches (North, 2000). Two prevalent issues have been put forward in creating scales for assessing oral performance (Brindley, 1998): (1) the number of levels and standards which are required and (2) the descriptor of each level must be clearly stated. Number of levels along with the number of standards is regarded as measurement-related issues whereas the definition of levels (descriptor) is concerned with description-related issues (North, 2000). Luoma (2004) believes that definition of each level must not be ambiguous, abstract and virtual. In addition, test makers require to discriminate the discourse appearing in the definition of the adjoining levels (Jin, Mak, & Zhou, 2012).

According to Luoma (2004), three approaches can be categorized for developing a sound rating scale for a particular speaking task i.e. intuitive, qualitative, and quantitative methods. Lately, two distinctive paradigms namely the measurement- driven and performance data-based methods were also suggested by Fulcher, Davidson, and Kemp (2011). In line with Luoma (2004), Poonpon (2009) asserts that there are quantitative methods with various research techniques which can be utilized to develop scales. Fulcher (2003) proposed that validity accounts should not only be considered on right after the development of the scale, but it should also be included in the initial stages of designing scales.

Using a quantitative approach, Wang (2015) claimed that verbal interaction features are clustered into four communication functions. He redefined the interactional competence operationally and argued that

interactional competence is "the ability to effectively talk with others in order to achieve different communicative goals such as building arguments, developing discussions, offering support and shaping connection among topics" (Wang, 2015, p.146). These communication functions are considered as the underlying factors predicting interactional competence in the CAP scale. As though, this study endeavors to investigate the psychometric characteristics of the CAP rating scale for interactional competence in paired-speaking tasks to examine empirical proof for the reliability and validity of the scale. This scale was developed by Wang (2015) to assess verbal interaction features of interactional competence at micro-level. Therefore the following research questions were raised:

RQ1: Is the CAP scale a reliable instrument to evaluate Iranian intermediate EFL learners' interactional competence?

RQ2: Is the CAP scale a valid instrument to evaluate Iranian intermediate EFL learners' interactional competence?

Method

Participants

The statistical population of the study were all intermediate EFL learners taking general English courses at a language institute in Tabriz, the center of East Azerbaijan Province, Iran. The initial sample consisted of 180 learners (102 females and 78 males). Their age ranged from 16 to 29 years. According to the principles of stratified random sampling, initially 14 intermediate classes out of 32 were randomly selected which included 180 learners. To ensure the homogeneity of the sample, a language proficiency test (TOEFL iBT) was administered prior to the study. Out of the initial sample, the learners whose score were one SD (Standard Deviation) above and one SD below mean were selected. Therefore, 20 students of the initial sample were omitted and there remained 160 students. The rationale behind the selection of intermediate learners was that intermediate learners are more likely to manifest a reasonable range of interaction features whereas due to the lack of proficiency, low proficient learners produce limited number of interactional features. Also, most advanced learners have already developed some interactional features hence, tracing the development of their interactional competence might be blurring.

It is worth mentioning that having run the statistical analysis, the scores of 36 students were dropped as being outliers. Finally, the actual data of 124 students were analyzed in the study. Interlocutors were paired to do the

interaction tasks and the pairings were the same during both pretest and posttest.

In line with previous studies (e.g. May, 2011) five experienced EFL teachers were employed in this study. They all acted as decision-makers who assigned scores to individual interlocutors' performance on each task. Raters were Language Institute teachers who were actively involved in assessing EFL learners' oral skills in placement test at the beginning of each term.

Instrumentation

The current study benefited two types of instrument: 4 paired-speaking tasks to provoke participants' interactional competence (see appendix), and the CAP scale as a measurement for the interactional competence at micro-level. The four tasks were distinctive based on the characteristics proposed by Ellis, (2003) and Samuda and Bygate, (2008) i.e., task outcome, Information access, and negotiation results.

As mentioned earlier, Wang (2015) argued that interaction features could be chunked through four communication functions rather than the two classifications suggested by Ducasse and Brown (2009). Filling a silence, making comments, dis/agreeing, back-channeling were subcategorized under the function of building arguments. Topic initiation, topic development, and topic connection encompassed the second function termed as developing discussions. The third function (offering support) was devoted to turns and the number of turns, turn interruption, and turn overlapping were its respected subordinate categories. And finally, confirmation question, opinion question, and information question were included in the forth communication function which is labeled as shaping connection.

Procedure

The study included three main phases. In the first phase, the researchers held two sessions for the raters in order to make them acquainted with the procedure of assigning scores in the adopted rating scale. All the raters/teachers had to attentively participate in the sessions. The researcher expounded the procedure of data collection for them and the raters' questions were addressed by the researchers. The four interaction tasks were presented and explained to the raters. They were told that the allowed time for each task is 2 minutes and a half so that there would be no need to normalize the scores. The procedure of pairing the student were clarified for them and it was highly recommended that the pairs have to be the same during both test and retest.

The second phase of the study was allotted to the first data collection session (testing session). The researchers coordinated the data collection date and scheduled the whole sections. All participants completed four tasks in pairs using their own cellphones. For the purpose of data-collection, in each session, two proctors were present. One of the researchers as the main proctor brought the sheets containing tasks to the session. Each sheet representing one task, was delivered to each participant in the specified pair. During the data collection, oral performance of the students for each task was audio-taped using students' mobile phones. The following steps were taken in the administration of the tasks:

1. Arranging pairs and assigning numbers to them;
2. Ensuring the availability of digital recorder for each pair;
3. Conforming the rubric of the task and completing it.

The third phase of the study was run similar to the previous phase with a fourteen week interval. The rationale behind this interval was to minimize the effect of learning from pre-test to post-test which could have confounded the results. The same tasks with the same procedure was administered to the students. Then the recording of each pair was transcribed by the raters and the scores were assigned to the individuals.

Results

Test-Retest Reliability Indices

The Pearson correlations were computed in order to probe the test-retest reliability indices of the 13 items related to four communication functions. The results are displayed in separate tables for each function.

Table 1 illustrates the test-retest reliability indices for building arguments.

Table 1
Pearson Correlations; Test-Retest Reliability Indices of Building Argument

		FSpost	MCpost	ADpost	BCpost
FSpre	Pearson Correlation	.671**			
	Sig. (2-tailed)	.000			
	N	124			
MCpre	Pearson Correlation		.906**		
	Sig. (2-tailed)		.000		
	N		124		
ADpre	Pearson Correlation			.874**	
	Sig. (2-tailed)			.000	
	N			124	
BCpre	Pearson Correlation				.866**
	Sig. (2-tailed)				.000
	N				124

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the results displayed in Table 1, there were significant relationships between the pretests and posttests of;

- filling the silence (FS) (r (122) = .671 indicating a large effect size, p = .000),
- making comment (MC) (r (122) = .906 indicating a large effect size, p = .000),
- agreeing/disagreeing (AD) (r (122) = .874 indicating a large effect size, p = .000), and
- back-channeling (BC) (r (122) = .866 indicating a large effect size, p = .000).

Table 2 indicates the Test-retest reliability indices for developing discussion.

Table 2
Pearson Correlations; Test-Retest Reliability Indices of Developing Discussion

		TIpost	TDpost	TCpost
TIpre	Pearson Correlation	.776**		
	Sig. (2-tailed)	.000		
	N	124		
TDpre	Pearson Correlation		.423**	
	Sig. (2-tailed)		.000	
	N		124	
TCpre	Pearson Correlation			.428**
	Sig. (2-tailed)			.000
	N			124

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the results displayed in Table 2, there were significant relationships between the pretests and posttests of;

- Topic initiation (TI) (r (122) = .776 indicating a large effect size, p = .000),
- Topic development (TI) (r (122) = .423 indicating a moderate effect size, p = .000), and
- Topic connection (TI) (r (122) = .428 indicating a moderate effect size, p = .000).

In table 3, the Test-retest reliability indices for offering support are provided.

Table 3

Pearson Correlations; Test-Retest Reliability Indices of Offering Support

		NTpost	TITpost	TOpost
NTpre	Pearson Correlation	.458**		
	Sig. (2-tailed)	.000		
	N	124		
TITpre	Pearson Correlation		.794**	
	Sig. (2-tailed)		.000	
	N		124	
TOpre	Pearson Correlation			.755**
	Sig. (2-tailed)			.000
	N			124

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the results displayed in Table 3, there were significant relationships between the pretests and posttests of;

- The number of turns (NT) (r (122) = .458 indicating a moderate effect size, p = .000),
- Turn interruption (Tit) (r (122) = .794 indicating a large effect size, p = .000), and
- Topic overlapping (TO) (r (122) = .755 indicating a large effect size, p = .000).

Table 4 shows the test-retest reliability indices for shaping connection.

Table 4

Pearson Correlations; Test-Retest Reliability Indices of Offering Support

		CQpost	OQpost	IQpost
CQpre	Pearson Correlation	.609**		
	Sig. (2-tailed)	.000		
	N	124		
OQpre	Pearson Correlation		.577**	
	Sig. (2-tailed)		.000	
	N		124	
IQpre	Pearson Correlation			.459**
	Sig. (2-tailed)			.000
	N			124

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the results displayed in Table 4, there were significant relationships between the pretests and posttests of:

- confirmation question (CQ) (r (122) = .609 indicating a large effect size, p = .000),
- opinion question (OQ) (r (122) = .577 indicating a large effect size, p = .000), and
- information question (IQ) (r (122) = .459 indicating a moderate effect size, p = .000).

Construct Validity of Communication Functions

Five separate exploratory factor analysis (EFA) through the varimax rotation method were conducted in order to examine the underlying constructs of the communication functions individually and as a total. To avoid repeating the same concepts, it should be mentioned that EFA has three main assumptions; sampling adequacy, lack of identity and lack of singularity which are tested using the KMO, Bartlett's chi-square and determinant statistics. If the KMO index is equal to or higher than .60 (Filed, 2013; Pallant, 2013), it can be concluded that the present sample size was adequate for running the EFA. If the Bartlett's test is significant ($p < .05$), it can be concluded that the correlation matrix is significantly different from an identity one; i.e. a correlation matrix with zero correlations among all variable. The opposite of identity is the singularity, a correlation matrix with perfect correlations among all variables. If the determinant value is higher than .00001, it can be concluded that the assumption of lack of singularity is met.

Construct validity of Building Argument.

An exploratory factor analysis was run to investigate the underlying constructs of the eight items of the building argument (Table 5). The assumptions were as follows:

- current sample size was not adequate for running the factor analysis (KMO = .347 < .60).
- the assumption of lack of identity was met (χ^2 (28) = 895.46, p = .000).
- the assumption of lack of singularity was met (Determinant = .001 > .00001).

Table 5
Total Variance Explained; Building Argument

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.037	25.462	25.462	2.037	25.462	25.462	1.925	24.064	24.064
2	2.018	25.229	50.691	2.018	25.229	50.691	1.911	23.892	47.956
3	1.806	22.575	73.266	1.806	22.575	73.266	1.889	23.614	71.570
4	1.565	19.558	92.825	1.565	19.558	92.825	1.700	21.255	92.825
5	.412	5.151	97.976						
6	.088	1.100	99.076						
7	.044	.547	99.622						
8	.030	.378	100.000						

The SPSS extracted four factors which accounted for 92.82 percent (Table 5) of the variance. In other words, the eight items related building argument measured four traits with an accuracy of 92.82 percent.

Table 6
Rotated Component Matrix; Building Argument

	Component			
	1	2	3	4
MCpost	.979			
MCpre	.970			
BCpost		.974		
BCpre		.950		
ADpost			.973	
ADpre			.959	
FSpost				.922
FSpre				.903

Table 6 displays the factor loadings of the eight items of the building argument under the four extracted factors. All factor loadings enjoyed large effect sizes (=> .50). Each pairs of items loaded on a distinct factor. That is to say, the pretest and posttest of making comments (MC) loaded under the

first factor. The pretest and posttest of back-channeling (BC) loaded under the second factor. The third factor included the pretest and posttest of agreeing/disagreeing (AD), and finally; the pretest and posttest of filling a silence (FS) loaded under the fourth factor.

Construct Validity of Developing Discussion.

An exploratory factor analysis was carried out to probe the underlying constructs of the six items of the developing discussion (Table 7). The assumptions were as follows:

- Present sample size was not adequate for running the factor analysis (KMO = .259 < .60).
- The assumption of lack of identity was met (χ^2 (15) = 826.24, p = .000).
- The assumption of lack of singularity was met (Determinant = .001 > .00001).

Table 7
Total Variance Explained; Developing discussion

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.293	38.221	38.221	2.293	38.221	38.221	2.119	35.319	35.319
2	1.659	27.647	65.868	1.659	27.647	65.868	1.830	30.497	65.816
3	1.312	21.866	87.734	1.312	21.866	87.734	1.315	21.918	87.734
4	.699	11.655	99.389						
5	.025	.411	99.800						
6	.012	.200	100.000						

The SPSS extracted three factors which accounted for 87.73 percent (Table 7) of the variance. In other words, the six items related developing discussion measured three traits with an accuracy of 87.73 percent.

Table 8 displays the factor loadings of the six items of the developing discussion under the three extracted factors.

Table 8
Rotated Component Matrix; Developing Discussion

	Component		
	1	2	3
TDpost	.940		
TCpost	.909		.355
TIpre		.932	
TIpost	.436	.894	
TCpre	.309		.751
TDpre	.334		-.739

All factor loadings in table 3, under their respective factors, enjoyed large effect sizes (=> .50). The pretest and posttest of topic development (TD) loaded under the first factor. The pretest and posttest of topic initiation (TI) loaded under the second factor, and finally; the pretest and posttest of topic connection (TC) loaded under the third factor.

Some of the items had minor loadings on other factors. For example; posttest of TI, and pretests of TC and TD had minor loadings under the first factor. The posttest of TC had a loading of .355 under the third factor. The loading of pretest of TD on the third factor was negative. That is to say; the third factor was a bipolar one. One of the variables had its loading on the negative side of the coordinate, while the other loaded on the positive side.

Construct Validity of Offering Support

An exploratory factor analysis was run to probe the underlying constructs of the six items of the offering support (Table 9). The assumptions were as follows:

- Present sample size was not adequate for running the factor analysis (KMO = .265 < .60).
- The assumption of lack of identity was met (χ^2 (15) = 1111.61, p = .000).
- The assumption of lack of singularity was met (Determinant = .0001 > .00001).

Table 9
Total Variance Explained; Offering Support

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.096	34.938	34.938	2.096	34.938	34.938	1.904	31.730	31.730
2	1.847	30.782	65.720	1.847	30.782	65.720	1.889	31.475	63.205
3	1.431	23.843	89.563	1.431	23.843	89.563	1.581	26.358	89.563
4	.614	10.232	99.795						
5	.009	.155	99.949						
6	.003	.051	100.000						

The SPSS extracted three factors which accounted for 89.56 percent (Table 9) of the variance. In other words, the six items related offering support measured three traits with an accuracy of 89.56 percent.

Table 10 displays the factor loadings of the six items of the offering support under the three extracted factors.

Table 10
Rotated Component Matrix; Offering Support

	Component		
	1	2	3
TITpost	.970		
TITpre	.903		
TOpost		.958	
TOpre		.891	
NTpost		.310	.868
NTpre			.820

All factor loadings in table 10, under their respective factors, enjoyed large effect sizes (=> .50). The pretest and posttest of turn interruption (Tit) loaded under the first factor. The pretest and posttest of turn overlapping (TO) loaded under the second factor, and finally; the pretest and posttest of number of turns (NT) loaded under the third factor. The posttest of number of turns had also a minor loading on the second factor.

Construct validity of Shaping Connection.

An exploratory factor analysis was run to probe the underlying constructs of the six items of the shaping connection (Table 11). The assumptions were as follows:

- Present sample size was not adequate for running the factor analysis (KMO = .268 < .60).
- The assumption of lack of identity was met (χ^2 (15) = 939.29, p = .000).
- The assumption of lack of singularity was met (Determinant = .0001 > .00001).
-

Table 11

Total Variance Explained; Shaping Connection

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.546	42.432	42.432	2.546	42.432	42.432	2.385	39.743	39.743
2	1.670	27.827	70.259	1.670	27.827	70.259	1.670	27.832	67.576
3	1.003	16.723	86.981	1.003	16.723	86.981	1.164	19.406	86.981
4	.740	12.333	99.314						
5	.038	.630	99.944						
6	.003	.056	100.000						

The SPSS extracted three factors which accounted for 86.98 percent (Table 11) of the variance. In other words, the six items related shaping connection measured three traits with an accuracy of 86.98 percent.

Table 12

Rotated Component Matrix; Shaping Connection

	Component		
	1	2	3
CQpost	.909	-.403	
OQpost	.842	.534	
IQpost	.842		.362
OQpre		.816	
CQpre	.312	-.746	
IQpre			.991

Table 12 displays the factor loadings of the six items of the shaping connection under the three extracted factors. Unlike the previous EFA

models, the present factor loadings did not show a clear pattern. The posttests of confirmation question (CQ), opinion question (OQ), and information question (IQ) loaded under the first factor. The pretests of confirmation question (CQ), and opinion question (OQ) loaded under the second factor; while the pretest of information question (IQ) alone loaded under the third factor. Some of the items had loading on other factors.

Construct Validity of Communication Functions

An exploratory factor analysis was run to probe the underlying constructs of the pretests and posttests of four communication functions (Table 13). The assumptions were the following:

- Present sample size was not adequate for running the factor analysis (KMO = .411 < .60).
- The assumption of lack of identity was met (χ^2 (28) = 1278.10, p = .000).
- The assumption of lack of singularity was met (Determinant = .0001 > .00001).

Table 13

Total Variance Explained; Communication Functions

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.399	42.487	42.487	3.399	42.487	42.487	3.127	39.086	39.086
2	1.534	19.175	61.662	1.534	19.175	61.662	1.569	19.607	58.693
3	1.166	14.570	76.231	1.166	14.570	76.231	1.349	16.860	75.553
4	1.052	13.147	89.379	1.052	13.147	89.379	1.106	13.825	89.379
5	.793	9.908	99.287						
6	.033	.410	99.697						
7	.015	.193	99.889						
8	.009	.111	100.000						

The SPSS extracted four factors which accounted for 89.37 (Table 13) percent of the variance. In other words, the eight components of communication functions measured four traits with an accuracy of 89.37 percent.

Table 14 displays the factor loadings of the eight components of the communication functions under the four extracted factors.

Table 14
Rotated Component Matrix; Communication functions

	Component			
	1	2	3	4
Post Shaping Connection	.963			
Post Developing Discussion	.946			
Post Offering Support	.927			
Pre Building Argument		.986		
Post Building Argument	.652	.735		
Pre Offering Support			.842	
Pre Shaping Connection			.732	
Pre Developing Discussion				.979

The posttests of shaping connection, developing discussion and offering support loaded on the first factor (Table 14). The pretest and posttest of building argument loaded under the second factor, while the latter had also a loading of .65 on the first factor. The pretests of offering support and shaping connection loaded under the third factor. The pretest of developing discussion loaded under the fourth factor.

Discussion

The accurate and sound assessment of all language skills in general and their sub-skills in particular, is an indispensable part of empirical research. Therefore, it is of paramount importance that we gauge and investigate the adequacy of our assessment instrument and procedures and weigh possible confounds in our teaching context.

What emerged strongly from data analysis and consequent results indicated that the test-retest reliability characteristics of the scale was quite satisfactory. It was found that there were significant relationships between the pretests and posttests scores of the 13 features of the scale ($p = .000$).

In terms of validity, having met the main assumptions of EFA i.e. sampling adequacy, lack of identity and lack of singularity, the four communication functions of building argument, developing discussion, offering support, and shaping connection as well as the overall communication functions measured the traits of their respected

subcategories with the accuracy of 92.82, 87.73, 89.56, 86.98, 89.37 percent respectively.

In terms of task type effect, it was found that the frequency of occurring some particular interaction features was more than others. For instance, opinion questions, turn connection, and agreement questions were more concerned with decision-making task indicating the openness and responsiveness of the interlocutors whereas turn interruption and turn overlapping, were the prominent interaction features appeared in story completion task . It can be inferred that story-completion task was more assertive and decisive task type. On the other hand, topic development, topic initiation, the number of turns, information questions, and confirmation questions features occurred more in spot-the-difference task showing a fractionally interactive conversation style in this type of task. Lastly, no feature was dominantly observed in free-discussion task since this type of task imposed no pressure on the interactants. Thus, it can be concluded that although the CAP scale accounted for all the interaction features, different task types provoke some interaction features more frequently than others.

The findings of the study are in line with Wang (2015) in which he found that the CAP scale is a reliable and valid measure for interactional competence. He found that the correlation coefficient for the 13 features of the scale ranged from .57 to .72 (Wang, 2015, p.124). In addition, through factor analysis he hypothesized that the interaction features are loaded into 4 factors namely argument, discussion, support and connection. The findings of the current study also support this claim. Therefore, it is in contrast with Ducasse and Brown's (2009) two factor model that verbal interaction features can be treated as two distinctive categories of interactional management and interactive listening

Finally, a note of caution appears to be necessary regarding the obtained scores of the study. The researchers approached an individualistic point of view for awarding scores for interaction performance. There are two legitimate concerns in this regard. It is argued that, on one hand, assigning shared scores in paired speaking tasks determines that the process of co-construction is reciprocal and a shared achievement (Norton, 2005; May, 2009). In this view the equity is ignored since the interlocutors might not donate to the conversation equally. On the other, designating individual scores for this kind of interaction may overlook the crucial reliance of the two interactants (McNamara, 1997). However, the procedure is complicated and future studies should be called upon to gather empirical evidence and

help test developers make sound decisions on how to award scores to interactional competence.

The findings of the study have several implications for language teaching in general and language testing in particular. Based on the findings of the study, teachers awareness with regard to interactional competence should be raised since operational definitions of the subcategories underlying this competence is validly and reliably established in the CAP scale considering the selected setting. Due to the significance of interaction which lies at the very core of interaction hypothesis (Long, 1981) and sociocultural theory (Ohta, 2000; Sun, 2012), teachers seem to obtain a better interpretation of the real manifestation of the competence in paired-speaking tasks; hence they can develop syllabi and accordingly tasks which can help the learners boost their interactional competence. As He and Young (1998) state, a language learner's interactional competence is considered local and practice specific. Hence, it can be enhanced through practice. In addition, teacher training programs can encompass courses for teachers and practitioners to make teachers familiar with the procedure of assigning scores to learners' performance and evaluate their interactional competence using this scale and other valid measures.

In the field of language testing, the most widely used tool in Iran's educational system is the format of multiple choice test. This device has been traditionally used since it is seen as being cheap, efficient, and reliable. However, it seems that there is a need for most Iranian teachers and practitioners to get familiar with various types of scale especially rating scales for assessing interactional competence in paired tasks since as mentioned earlier, there has been increasing demand for paired speaking assessment (Galaczi, 2013) and the frequency of integrating paired-speaking tasks in commercially published textbooks for example Oxford University Press and Cambridge University Press is expanding.

Nevertheless, the researchers acknowledge that the scope of application of the results of this study is restricted to the reliability and validity accounts of the CAP rating scale for assessing Iranian intermediate EFL learners' interactional competence using the four interaction tasks. Generalizing the findings from the present study to other contexts should be done with caution given possible differences in terms of L1 background, socio-cultural norms, and educational backgrounds. Some research in other contexts and with various participants is required to collect empirical evidence to consolidate the results of the research.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Brindley, G. (1998). Describing language development: Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interface between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.

Bachman & A. D. Cohen (Eds.), *Interface between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.

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*, 1-47.

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.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Filed, A. (2013). *Discovering Statistics Using IBM SPSS, Statistics for Statistics*. (4th ed.). London: SAGE Publications.

Fulcher, G. (2003). *Testing second language speaking*: Pearson Education.

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

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29.

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

Galaczi, E. D. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553–574.

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

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. *Talking and testing: Discourse approaches to the assessment of oral proficiency*, *14*, 1–24.

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

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

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing*, *29*(1), 43–65.

Kley, K. (2015). *Interactional competence in paired speaking tests: role of paired task and test-taker speaking ability in co-constructed discourse* (Unpublished doctoral dissertation). University of Iowa, Iowa.

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

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book*. New York: McGraw-Hill Book.

Long, M. H. (1981). Input, interaction, and second-language acquisition. In H. Winitz (Ed.), *Native Language and Foreign Language Acquisition*,

(pp. 259-278). New York: Annals of the New York academy of sciences.

Luoma, S. (2004). *Assessing speaking*. 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.

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

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*(4), 445-465.

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

Ohta, A. S. (2000). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In J. P. Lantolf *(Ed), Sociocultural theory and second language learning*, (pp. 51–78). Oxford: Oxford University Press.

Oksaar, E. (1990). Language contact and culture contact: Towards an integrative approach in second language acquisition research. *Current Trends in European Second Language Acquisition Research. Multilingual Matters, Clevendon*, 10–20.

Pallant, J. (2013). *SPSS survival manual*. London: McGraw-Hill Education.

Poonpon, K. (2009). *Expanding a second language speaking rating scale for instructional and assessment purposes*. Arizona: Northern Arizona University Press.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke: Palgrave Macmillan.

Sun, Y. (2012). The influence of the social interactional context on test performance: A sociocultural view. *Canadian Journal of Applied*

Linguistics/Revue canadienne de linguistique appliquée, *14*(1), 194–221.

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

Taylor, L., & Wigglesworth, G. (2009). *Are two heads better than one? Pair work in L2 assessment contexts*: Sage Publications Sage UK: London, England.

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

Appendix A-1 Interaction Tasks

Task 1: Spot Differences

(Student A)

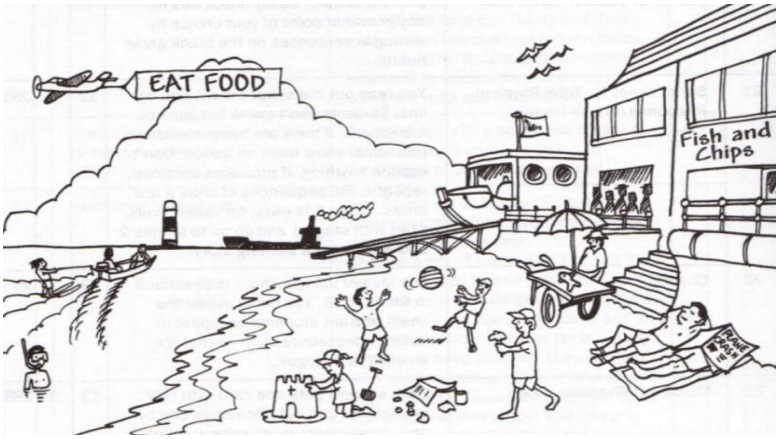
Directions: Look at the picture below. **Don't look at your partner's picture**!

Your partner has a similar picture with some minor differences. There are over 10 differences between the two pictures.

Discuss with your partner in **English** to:

Find <u>at least 5 differences</u> in the two pictures by asking and answering questions.

You have <u>1 minute</u> to prepare and <u>2.5 minutes</u> to record your conversation.



(Source of the pictures: http://mayitaclass.blogspot.com/). Adopted from (Wang, 2015)

Task 1: Spot Differences (continued)

(Student B)

Directions: Look at the picture below. **Don't look at your partner's picture**.

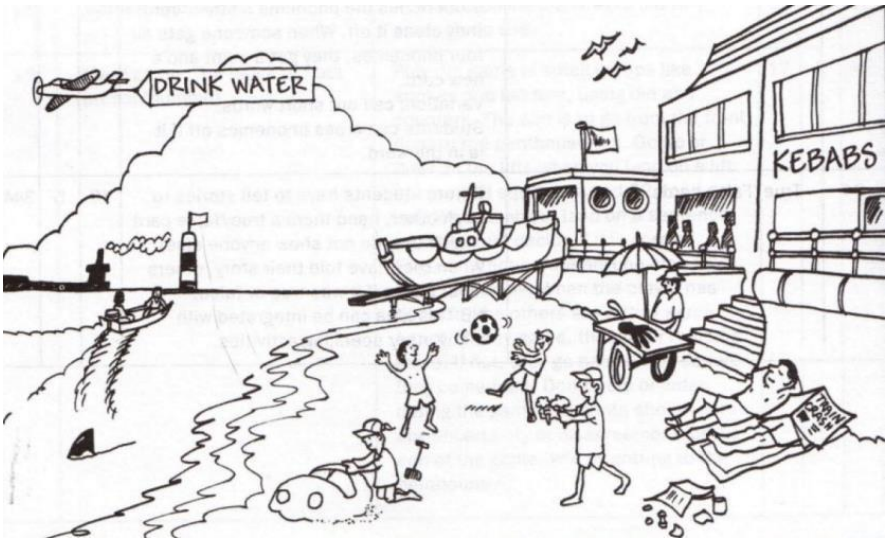
Your partner has a similar picture with some minor differences. There are over 10 differences between the two pictures.

Talk to your partner in **English** to:

Find <u>at least 5 differences</u> in the two pictures by asking and answering questions.

You have <u>1 minute</u> to prepare and <u>2.5 minutes</u> to record your conversation.

Picture B



Appendix A-2

Task 2: Complete a Story

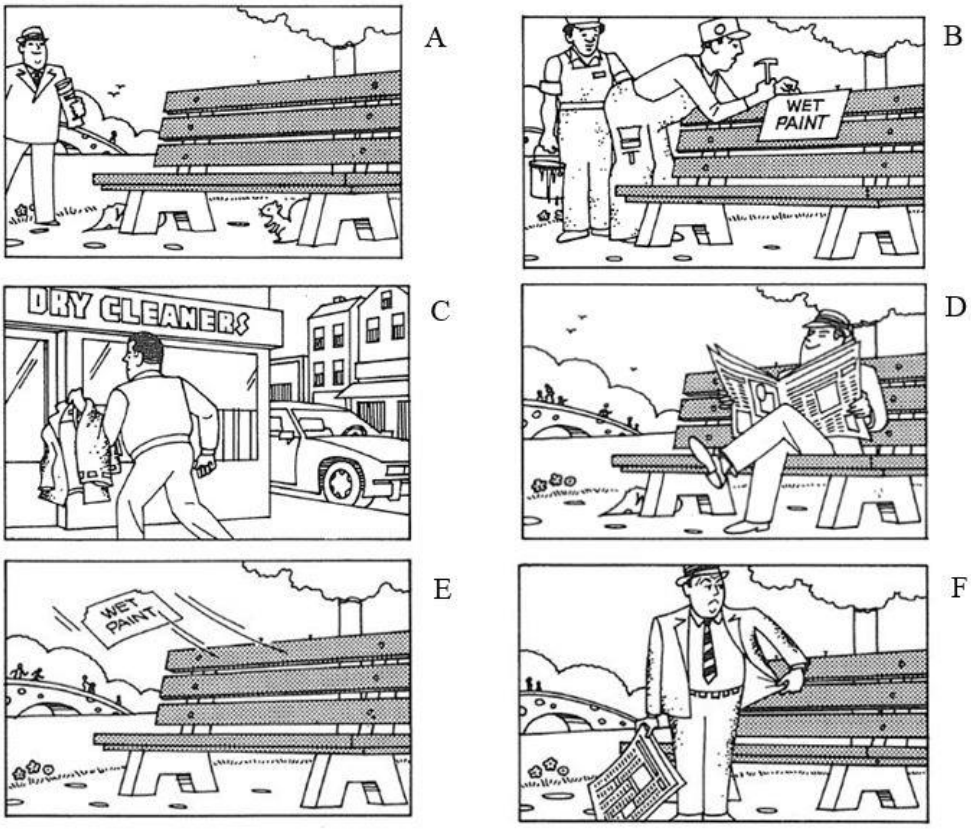
Directions: You and your partner have a set of six pictures which tell a complete story. But the pictures are NOT in the right order now.

Talk to your partner in **English** to:

Arrange the pictures <u>in the right order</u> to tell the complete story

Refer to each picture by its letter (A, B, C, D, E, and F).

You have <u>1 minute</u> to prepare and <u>2.5 minutes</u> to record your conversation.



(Adapted from SPEAK Practice Test: Wet Paint Sign). Adopted from (Wang, 2015)

Appendix A-3

Task 3: Decision-making

Directions: You and your partner are going to discuss the topic of a presentation you must give in class called:

"Our Favorite Healthy Food"

Talk to your partner **in English** to:

1. Choose what food you think should be in the presentation and why;
2. Reach an agreement on which food you are going to talk about in your presentation.

You have 1 minute to prepare and 2.5 minutes to record your conversation.

Adopted from (Wang, 2015)

Biodata

Milad Ramazani is an assistant Professor at IAU, Urmia Branch. He has trained BA students and supervised MA theses in ELT. He has also instructed IELTS and TOEFL courses since 2008. He has presented in some national and international conferences. His research interests include Psycholinguistics, Bilingualism/Multilingualism, Conversation Analysis, Interactional Competence, and Materials Evaluation.

Biook Behnam is an associate Professor of Applied Linguistics at Islamic Azad University, Tabriz branch, Iran. His current research interests cover Discourse Analysis, ELT and Translation Studies. He has been involved in a wide range of projects in the area of Applied Linguistics and Discourse Analysis as a project director, consultant and researcher.

Saeideh Ahangari is an assistant professor in TEFL at Islamic Azad University/ Tabriz Branch. Her main interests are task-based language teaching, CALL and their interface with the issues in language testing. She has published many articles and participated in many national and international conferences.