Journal of Language Testing & Assessment (2019) Vol. 2: 1-12
Clausius Scientific Press, Canada

DOI: 10.23977/langta.2019.21001
ISSN 2616-2938

INTERACTION FEATURES IN PREDICTING COMMON EUROPEAN FRAMEWORK OF REFERENCE LEVEL

Okim Kang[1,a,*], Garrett M Larson[1,b] and Soo-Hyun Koo[2,c]

[1]Northern Arizona University, Flagstaff, Arizona, USA
[2]Department of English Education, Seoul National University, Seoul, Republic of Korea
a. Okim.Kang@nau.edu, b. gl349@nau.edu, c. koosam@snu.ac.kr
**corresponding author*

Keywords: Paired speaking Assessment, Proficiency Level, Interaction Features, Canonical Discriminant Analysis

Abstract: The study examined how interaction features could predict test takers' English proficiency levels in high-stakes contexts. Using candidates' interactive task responses in the Cambridge English Language Assessment (CELA), it explored salient interaction features that could distinguish across Common European Framework of References (CEFR) speaking levels (B1- C2). It further ascertained the degree of accuracy in the proficiency prediction through a canonical discriminant analysis (CDA) in which weighted contributions of the 11 interaction features were created to best predict CEFR level of each test-taker. Fifty-eight video files (i.e., 116 interactive speech samples) were coded for the categories of four interactive features: (1) co-operation, (2) coherence, (3) turn-taking, and (4) strategy use. The results suggest that the selected interaction features distinguish test-takers' CEFR levels with over 50-60 % accuracy especially regarding interactive turn and initiation. Findings offer direct implications to ESL classrooms and provide evidence to enhance our understanding of the complex nature of interactional competence in the context of the high-stakes speaking assessment.

1. Introduction

Communication is a two-way street. Successful communication happens only when both interlocutors are actively engaged. That is, both active listeners and eloquent speakers should be responsible for their communication (Kang, Rubin, & Lindemann, 2015). Such communication can be ensured by effective and well-managed interaction. Interaction is a dynamitic process of communication. It is often described through interactional competence (IC), which can be referred to as interlocutors' ability to accomplish meaningful social goals and to make their actions recognizable to others (Hall & Pekarek, 2011). In this study, we attempted to examine a wide range of interaction features used in roles of the speaker and the listener to co-construct successful communication in order to achieve a higher score in the oral assessment context.

In the context of oral assessment, the dynamic processes between interlocutors take place as test-takers work together to achieve their communicative goals through their talk-in-action. Each interlocutor's contribution to their interaction can often be recognized by scores from each interlocutor in a speaking task based on the use of interaction features. Interactional features can be measured to illustrate and quantify the IC. It is not surprising to assume that interlocutors may use interaction features differently depending on the level of their IC. IC acknowledges the collaborative and co-constructed nature of conversations (He & Young, 1998; Kramsch, 1986). As interlocutors use various interaction features either to signal their listenership (a.k.a. interactive listening) or their understanding of interlocutors' message (a.k.a. interactional management), a speakers' IC is mostly demonstrated through timely appropriate use of interaction features during the conversation.

Noting IC's importance in oral communication, several studies (Brooks, 2009; Ducasse & Brown, 2009; Galaczi, 2008) investigated the IC of interlocutors conversing in their second languages (L2). They investigated how interlocutors with varying levels of L2 proficiency use interaction features differently. However, little has been done regarding whether IC components can accurately predict L2 proficiency. In particular, it is largely unknown how interaction features are associated with high-stakes test scores. Accordingly, the current study investigated to what extent interaction features could predict L2 English speakers' proficiency level. Using candidates' interactive task responses in the Cambridge English Language Assessment (CELA), the study identified salient interaction features that could distinguish across Common European Framework of References (CEFR) speaking levels (B1- C2). It further ascertained the degree of accuracy in the prediction process.

2. Literature review

2.1. Role of Interaction Features in L2 oral Communication

Interaction features play a primary role in L2 oral communication, specifically in the concept of IC. IC can be referred to as a dynamic process of communication built through the collaborative effort of the interactional partners (Kramsch, 1986). The collaborative effort of both interlocutors to co-construct conversations has been considered a key characteristic of successful interaction (He & Young, 1998). The degree of co-construction can be viewed at both the macro level of overall interaction quality and the micro level of interaction features. The overall quality of co-construction is connected to the equality of power and mutuality. The concepts of mutuality and equality can illustrate different levels or quality of interaction. As Oksaar (1990) states, interaction features represent the ability of a person to carry out and interpret verbal, paralinguistic, non-verbal, and extraverbal communicative actions in both the speaker's and the hearer's roles. IC is frequently described through verbal features like taking turns in a talk and repairing problems (Kasper, 2006). The current study defines the concept of IC as a co-constructive process in roles of the speaker and listener needed for successful communication. This study also hypothesizes that IC can be quantitatively measured through a range of interaction features.

Perhaps, one of the most well-known studies about interaction features would be Storch (2001, 2002), even though the patterns were described for ESL writing papers: collaborative, dominant/passive, expert/novice, and dominant/dominant. Dimitrova-Galaczi (2004) furthered the idea of interaction features for a paired speaking task on peer-peer interaction on the First

Certificate in English. She argues that collaborative interactions have a moderate to high degree of mutuality and equality. Collaborative patterns are considered as the best type of interaction, while dominant interlocutors are characterized as having low mutuality since they normally focus more on leading the conversation than intentionally involving with the passive one. On the other hand, passive ones tend to follow and react to the dominant ones. Overall, interaction patterns are strongly related to the overall quality of interaction. A high level of mutuality is valued when evaluating interaction quality.

2.2. Relationship between Interaction Features and Common European Framework of Reference (CEFR) level Prediction

Previous studies not only categorized the components of interaction features, but they also found the relationship between interaction features and L2 learners' proficiency. They found that speakers with high L2 proficiency showed (a) equal number of turn initiations (Galaczi, 2013), (b) frequent and adequate use of back-channelling and prompting (Ducasse & Brown, 2009; May, 2011), and (c) variety of question type use (Brooks, 2009). On the contrary, speakers with low L2 proficiency showed (a) frequent turn interruption, (b) frequent overlapping initiation (Ducasse & Brown, 2009; Dimitrova-Galaczi, 2004). However, findings about certain interaction features have not always come to a clean-cut argument between researchers. For instance, while a higher score was given to learners who demonstrated a range of speech functions (i.e., agreement, explanations, challenges) in Gan's (2010) study about classroom assessment, Brooks (2009) found that higher levels of interaction (i.e., follow-up questions, requests of clarification) did not guarantee higher scores because of the presence of negatively perceived interaction features, such as turn overlaps and misunderstanding. Such signs of disagreement call for the need of a study that not only examines the validity of predicting IC, but also one that uses an alternative approach.

Despite the contribution that previous studies have made to broaden the understanding of IC and its relationship with CEFR proficiency, some limits still exist. Most of the previous studies tried to identify the characteristics of interaction features of language learners after their proficiency is already known. In other words, little is known about the matchedness of predicted proficiency based on interaction feature use, and the actual proficiency of L2 language learners. In addition, we still do not know whether interactional features are good predictors across proficiency (e.g., beginner to advanced), or just for specific level of proficiency. There is a need for a study to examine the validity of the predictive power of these features. For the connective nature of proficiency levels between the Cambridge English exam's level and CEFR, the study used these terms somewhat interchangeably. The following research questions has guided the study.

1. What are the interaction features that distinguish the Cambridge English exams' levels (CEFR B1-C2)?
2. How accurately can interaction features predict the Cambridge English exams' levels (CEFR B1-C2) of test-takers?

3. Method

3.1. Speech Samples

The study used 58 video files from an interactive task, which were provided by the CELA. It is one of the largest international standardized English tests and employs a series of tests that reflect the proficiency levels B1 to C2 of the CEFR. These tests include the Preliminary English Test (PET: B1), the First Certificate in English (FCE: B2), Certificate in Advanced English (CAE: C1), and

Certificate of Proficiency (CPE: C2), which are equivalent to the CEFR levels B1, B2, C1, and C2, respectively. The CELA only provided speech samples from test-takers who passed their respective test, which enables us to believe that each test represents each interlocutors' actual English language proficiency level. This includes 14 videos from the B1 group, 16 from the B2 group, 17 from the C1 group, and 11 from the C2 group. Two minutes from each video file were extracted for a total of 116 minutes of speech. Among the candidates, 74 were females and 42 were males. Participants were from many different L1 backgrounds (e.g., Spanish, Korean, Italian, Dutch, French, Chinese, Japanese, German, Portuguese, and Russian).

3.2. Data Coding

The 116 minutes of interactive speech was first transcribed. The term, *interactive communication*, in Cambridge English exams, is a construct used to measure test-takers' communicative competence. Interaction skills are assessed by criteria such as initiating and responding, contributing to conversation development actively, and also using interactive strategies to maintain and repair communication. In order to operationalize IC and to determine the variables, Celce-Murcia, Dörnyei, and Thurrell's (1995) discourse and strategic competence framework was used. The transcribed speech was then manually coded for interaction features, which included the following subsections: (1) co-operation, (2) coherence, (3) turn-taking, and (4) strategy use. Table 1 provides a list of the specific interaction features (k = 11) that represent each subsection. Inter-coder reliability was determined to be acceptable, as it was above 0.75.

Table 1 Summary of interaction features analysed.

Criteria	Sub-measures
Co-operation	Back-channelling
	Prompting
	Topic initiation
	Overlapping initiation
Coherence	Discourse markers
Turn-taking	No. of total turns
	No. of short turns (1-10 words)
	No. of middle turns (11-30 words)
	No. of long turns (more than 30 words)
Strategy use	Response maintenance
	Repair

Coperation measures were operationalized by variables of back-channeling (the occurrence of utterances briefly responding to a partner in either nonverbal units or phrasal ones), prompting, topic initiation (the occurrence of utterances starting a new idea), and overlap initiation (the occurrence of two interlocutors starting their utterances at the same time) in conversations employing a conversation analysis approach (Schegloff, 1982). The measure of coherence included discourse markers. They are words or phrases that are relatively syntax-independent and do not change the meaning of the sentence. When it comes to the turning-taking measures, using Crookes's (1990) definition of a "turn", we operationalized it as one or more streams of speech bounded by speech of another, usually an interlocutor (Crooke, 1990). Regarding turn-lengths, the study adopted Gnisci and Bakeman's methods (2007) to measure turn lengths as short (1-10 words), middle (11-30 words), and long (more than 30 words). Finally, with regard to the measure of strategy use, we used Nakatani's (2010) response maintenance (utterances mentions a part of previous utterances of other interlocutors) and repair strategies (utterances corrects one interlocutor's own speech).

In sum, the interaction measures included in the current study were the following 11 features: back-channeling, prompting, topic initiation, overlapping initiation, discourse-markers, total talking time, turn-taking time, number of total turns, total number of short turns (one to ten words, total number of middle turns (eleven to 30 words), total number of long turns (more than 30 words), response maintenance (utterances of referring to the other interlocutor's previous utterances), and repair (utterances of self-correcting).

3.3. Data Analysis

The study employed a canonical discriminant analysis (CDA) to examine how well interaction features can predict CEFR level. CDA is interested in the predictive power of combined variables in determining previously set levels, which is unlike multiple regression analyses which are interested in the predictive power of each individual variable. By reducing the number of variables into functions, CDA linearly combines observable phenomena into new latent (unobservable) variables (Plonsky, 2015). Discriminant analyses reverse the independent and dependent variables around in a traditional MANOVA. Using the variables in this study, in a traditional MANOVA, we might examine proficiency level (IV) and how it affects interaction feature use (DV). In a discriminant analysis, the IVs and DVs are switched. Thus, the current study examined how well interaction feature use (IV) can predict proficiency level (DV). Using SPSS, all statistical assumptions were checked before running the final analysis. Concerning normality, even though some features failed to retain the null hypothesis, they were kept in the analysis after reviewing their respective histograms and Normal Q-Q plots. Correlations among the features appeared to be linear, and there is an absence of multivariate outliers. Even though one observation exceeded the critical value for Mahalanobis Distances, it was kept because the difference was not great (Tabachnick & Fidell, 2007). Although Box's M test was significant ($p = 0.000$) and there appears to be multicollinearity, based on Tolerance and VIF values outside of the acceptable range, all features were kept for the final analysis because it produced a model that was best at predicting CEFR level. In sum, the data analysis created weighted contributions of the 11 interaction features that could best predict the proficiency or CEFR level of each test-taker.

4. Results

The study attempted to examine to what extent interaction features could distinguish the CEFR B1-C2 levels by using a CDA. The analysis first revealed relationships between interaction features and CEFR level by using descriptive statistics (i.e., means and standard deviations). Overall, the 11 interaction features, which represent CEFR level in terms of feature use frequency, significantly distinguish among the four groups at the 0.05 level: Wilks' $\Lambda = 0.373$, $X\ 2\ (30) = 106.60$, $p = 0.000$. Figure 1 provides a visual representation of Table 2, which show the relationships between CEFR level and interaction feature use frequency.

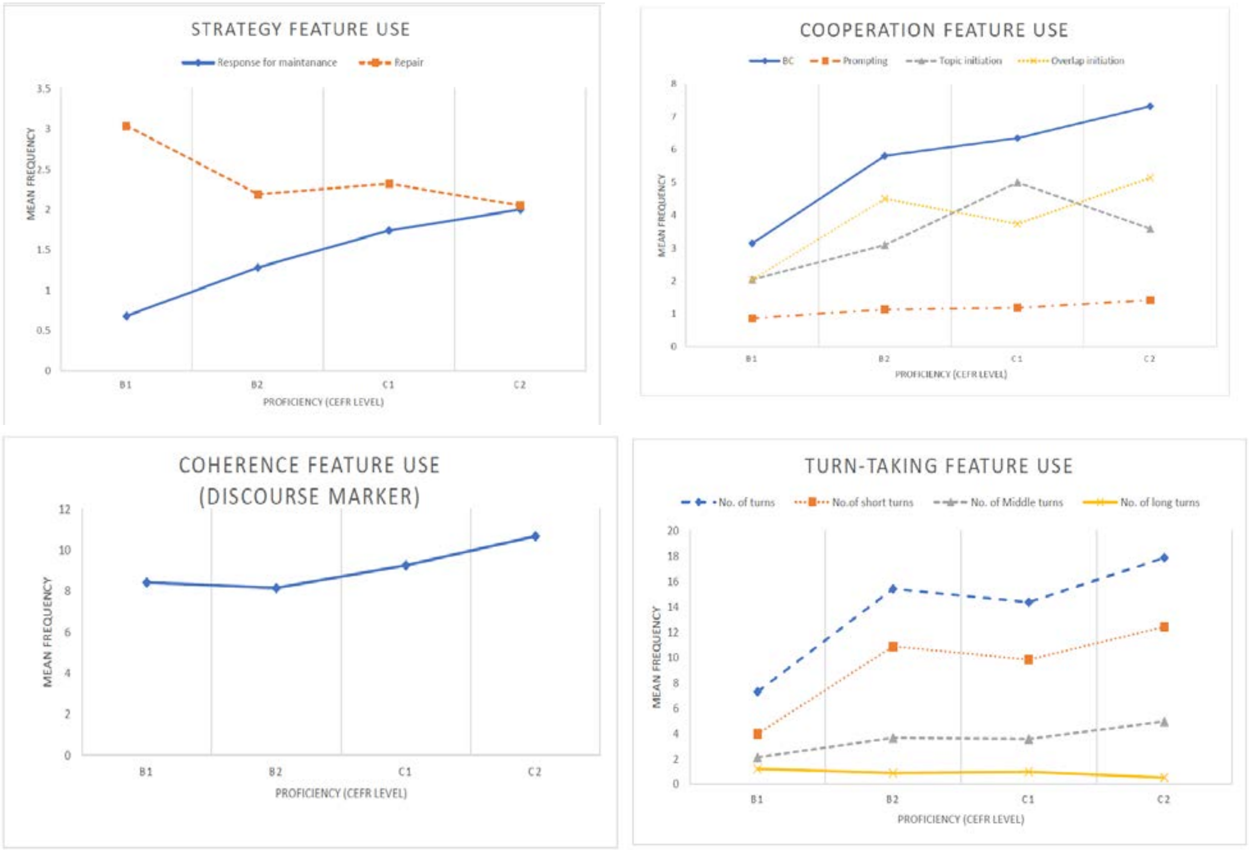

Figure 1 Mean frequency of interaction features according to CEFR level.

As we can see, the first noticeable trend is that, overall, the frequencies of interaction features are linearly related with CEFR level. We can see this relationship in the use of back-channelling, overlapping initiations, number of turns, number of short/middle/long turns, and response maintenance. In other words, as proficiency increases from B1 to C2, these features are generally increasing as well. Even though a linear trend is also shown for prompting, discourse markers, and repair, the *p*-value is not significant, which means that the differences between each group are not significant for these features. Topic initiation is the only interaction features which shows a decrease in mean frequency as CEFR level increases from C1 to C2.

Table 2 Mean frequency counts of interaction features.

Variable	Group statistics				Tests of equality of Group means		
	PET (B1)	FCE (B2)	CAE (C1)	CPE (C2)			
	Mean frequency (SD)	Mean frequency (SD)	Mean frequency (SD)	Mean frequency (SD)	Wilks' Λ	*F*	*P*
Backchannelling	3.14 (3.00)	5.81 (3.10)	6.35 (4.97)	7.32 (4.40)	.87	5.40	.002*
Prompting	.86 (.81)	1.13 (1.07)	1.18 (1.00)	1.41 (1.44)	.97	1.12	.345
Topic initiation	2.04 (.96)	3.09 (1.49)	5.00 (2.26)	3.59 (1.62)	.69	16.72	.000**
Overlap initiation	2.04 (1.80)	4.50 (3.81)	3.74 (3.21)	5.14 (2.90)	.88	5.04	.003*
Discourse Marker	8.43 (3.72)	8.16 (3.43)	9.26 (4.41)	10.68 (4.10)	.95	2.09	.106

No. of turns	7.29 (4.38)	15.41 (6.33)	14.35 (7.10)	17.86 (5.33)	.71	15.24	.000**
No. of short turns	3.96 (3.87)	10.88 (6.35)	9.82 (6.51)	12.41 (5.50)	.77	11.11	.000**
No. of middle turns	2.11 (1.60)	3.66 (1.98)	3.56 (2.05)	4.95 (2.23)	.81	8.77	.000**
No. of long turns	1.21 (.79)	.88 (.707)	.97 (1.00)	.50 (.80)	.92	3.06	.031*
Response for maintenance	.68 (.77)	1.28 (1.25)	1.74 (1.42)	2.00 (1.41)	.87	5.58	.001**
Repair	3.04 (2.67)	2.19 (1.69)	2.32 (2.00)	2.05 (2.10)	.97	1.16	.329

Note. **p* < .05, ** *p* <.001

The 11 interaction features were reduced into functions. The current analysis categorized the 11 features to three functions (Interactive Turn and Initiation, Discourse Markers, Absence of Turns). Below, Table 3 shows how strongly each interaction feature correlates with each function. The asterisk marks the largest absolute correlation between each feature and its respective function.

Table 3 Structure matrix of canonical discriminant analysis.

Feature	Function		
	1	2	3
Topic initiation	.778*	.405	.127
No. of turns	.710*	-.465	.068
No. of short turns	.608*	-.394	-.019
No. of middle turns	.487*	-.387	.387
Response maintenance	.456*	-.083	.376
Back-channelling	.456*	-.171	.180
Overlap initiation	.364*	-.341	.010
Repair	-.197*	.125	.024
Prompting	.184*	-.107	.169
Discourse marker	.149	-.058	.544*
No. of long turns	-.233	.284	-.287*

As we can see in Table 3, nine interaction features correlated strongest with Function 1. However, only three are strongly correlated (i.e., above 0.50) with Function 1: topic initiation (0.778), number of turns (0.710), and number of short turns (0.608). This means that, according to Function 1, interlocutors’ CEFR level is influenced mostly by co-operation (i.e., topic initiation) and turn-taking features (i.e., no. of turns and no. of short turns). None of the 11 interaction features correlated most with Function 2. Discourse markers correlated strongest with Function 3 (0.544). Even though number of long turns correlated most with Function 3, the strength of the correlation was very weak (-0.287).

After each of the features loading heaviest on each function were examined, Function 1 was labelled as “*Interactive Turn and Initiation*”, while Function 3 was labelled as “*Discourse Markers*”. Even though none of the 11 features correlated most with Function 2, we labelled it “*Absence of Turns*”, based on the negative correlations with the features number of turns and number of short and middle turns.

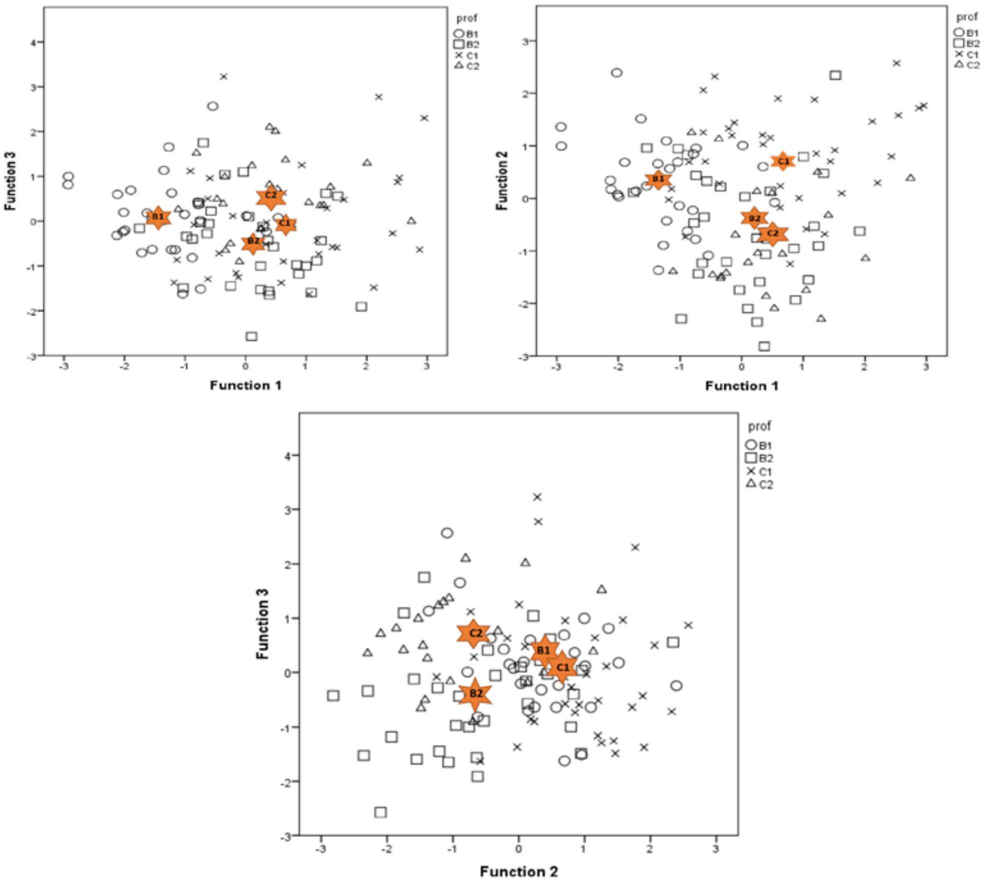

Figure 2 Scatter plots for group centroids.

In Figures 2-1 to 2-3, the orange stars are identified with one of the four CEFR levels: B1, B2, C1, and C2. Each star represents the group centroid. As we can see in Figures 2-1 and 2-2, “Interactive Turn and Initiation” (Function 1) can distinguish B1 test-takers from the other groups quite well, but not among the B2, C1, and C2 groups since their centroids are close together. “Absence of Turns” (Function 2) seems to be able to distinguish between B1/C1 and B2/C2 test-takers well based on the proximity of their group centroids in Figures 2-1 and 2-3. However, it cannot distinguish between B1 and C1 test-takers, or between B2 and C2 test-takers. Lastly, looking at Figures 2-2 and 2-3, “Discourse Markers” (Function 3) can distinguish between B2 and C2 test-takers quite well, but not B1 and C1.

Accuracy measures (i.e., precision and recall) are given in Table 4 below. Attention is given to the cross-validated values because they are more conservative and robust values. Precision and recall are important to consider because they tell us how well the current model can guard against type I and type II errors (i.e., false positives and false negatives, respectively). Precision is the number of correct predictions divided by the total number of predictions the model makes. The precision value for B1 test-takers is 0.53. This means that of all the test-takers that were predicted to be in level B1, 53% of those predictions were correct. Recall is the number of correct predictions divided by all the observations that should be classified as such. Looking at the cross-validated recall values, we can see that the value is 0.68 for level B1, i.e., 68% of the B1 test-takers were classified accurately. Both the original and cross-validated values are given in Table 4.

Table 4 Accuracy of the model in terms of precision and recall rate.

Level	Original		Cross-validated	
	Precision	Recall	Precision	Recall
PET (B1)	0.60	0.75	0.53	0.68
FCE (B2)	0.61	0.53	0.57	0.41
CAE (C1)	0.71	0.59	0.62	0.53
CPE (C2)	0.48	0.55	0.39	0.50
Mean	0.60	0.61	0.53	0.53

Accordingly, the current model has an overall precision and recall rate of 53%. The 11 interaction features are most accurate in predicting C1 group membership since the precision value is highest for this CEFR level (0.62). This means that of all the test-takers that the current model predicted to be in level C1, 62% of those predictions were correct. In terms of the recall values, the interaction features are most accurate in predicting B1 test-takers since the recall value is the highest for this CEFR level (0.68); that is, 68% of the B1 test-takers were predicted to be B1. The 11 interaction features are least able to predict C2 test-takers since the precision value is lowest for this CEFR level (0.39), meaning that only 39% of the test-takers the model predicted to be C2 were correct. Looking at recall, the model is least able to predict B2 test-takers (0.41) as only 41% of the B2 test-takers were accurately placed into that group.

Table 5 Classification results

		Predicted Group Membership				Total
		B1	B2	C1	C2	
Original	B1	21	3	2	2	28
	B2	8	17	3	4	32
	C1	5	2	20	7	34
	C2	1	6	3	12	22
Cross-validated	B1	19	3	2	4	28
	B2	9	13	5	5	32
	C1	6	2	18	8	34
	C2	2	5	4	11	22

Table 5 looks more closely into the findings of precision/recall rate, and shows the dispersion of mismatched proficiency cases. For instance, in original dataset, 6 B1 level interlocuters were actually predicted to be appropriate for other levels. Out of those 6 B1 interlocuters, 2 cases were predicted to be B2, 3 to be C1 and 1 to be C2. The further the distance between predicted CEFR level and actual CEFR level shows how prediction can be misleading and whether that degree of misleading is serious.

5. Discussion

To reiterate, the current study aimed to answer the following research questions: (1) What are the interaction features that distinguish the Cambridge English exams' levels (CEFR B1-C2)? And how accurately can interaction features predict the Cambridge English exams' levels (CEFR B1-C2) of test-takers? The overall findings of the study suggest that interaction features can predict CEFR levels with a precision and recall value of 0.53. For precision, this means that 53% of the predictions that the model makes are correct. Recall rate represents that out of all the test-takers that should be classified in level X, 53% were accurately placed. The CDA findings suggest that the interaction features can be clustered into three distinct functions, related to the following: (1) interactive turn and initiation, (2) discourse markers, and (3) absence of turns. Using the three functions borne out of the CDA, we can distinguish between clusters of CEFR levels but not each individual level itself. Function 1 (i.e., interactive turn and initiation) can distinguish B1 test-takers from the other groups, Function 2 (i.e., absence of turns) can to distinguish B1/C1 test-takers from C2/B2 test-takers, and Function 3 (i.e., discourse markers) can only distinguish B2 and C2 test-takers.

As we can see in the scatter plots (Figure 2), Function 1 (i.e., interactive turn and initiation) can distinguish B1 test-takers from the other groups in that interaction competence is a dynamitic process of co-construct (He & Young, 1998). However, it cannot distinguish among the B2, C1, and C2 groups. This suggests that the features associated with Function 1 can isolate B1 speakers from all other test-takers. The reason Function 1 cannot distinguish among the higher CEFR levels might be explained by looking at the relationships between the interaction features in Figure 1. Even though the patterns are linear, the total number of turns, number of short turns, and overlap initiation all spike downwards at the C1 level; meaning that as CEFR level increases, these features rise but dip at the C1 level and then rise again. This corresponds with a spike and subsequent fall in topic initiations at the C1 level to the C2 level. This suggests an interaction effect among the Function 1 features which may have made it more difficult for the model to detect differences among the B2, C1, and C2 levels. To explain this, it is possible that more topic initiations in an interactive task come at the expense of the total number of turns and overlap initiations. This suggests that the more learners initiate new topics, they produce fewer turns and overlaps.

This explanation can also help us understand Function 2 (i.e., the absence of turns). As the results show, Function 2 can distinguish between the B1/C1 levels as a group from the B2/C2 levels. This suggests that as learners produce fewer turns, the model predicts that they will be in either level B1 or C1. A look at the relationships can again help explain this finding. Looking at Figure 1, we can see that from B1 to B2, the total number of turns increases. From B2 to C1, it decreases, and from C1 to C2 it increases again. The B2 group produced more turns in total and more short/middle turns than the C1 group. This explains why the model can use these features to distinguish the B1/C1 levels from the B2/C2 levels. Similar to Function 1, this pattern can again be explained by the C1 group’s spike in topic initiation use which comes at the expense of the total number of turns, as well as short and middle turns.

Moving on to Function 3 (i.e., discourse markers), the CDA findings show that this feature can distinguish between B2 and C2 test-takers but not B1 and C1 test-takers. This is not surprising considering the relationship among the levels we see in Figure 1. As we can see in Figure 1, discourse marker use decreases slightly from B1 to B2, and then rises at each subsequent level. However, these differences were not statistically significant. The only significant difference seems to be between the B2 and C2 levels which we can see in the line graph. This may explain why Function 3 can distinguish between the upper B and C levels but not the lower ones.

Concerning the accuracy of the overall predictions, the findings show that both the precision and recall values for the cross-validated results have a value of 0.53. Looking at precision, the model was most precise when predicting C1 test-takers (i.e., 0.62). This means that of the test-takers predicted to be in level C1, 62% of those predictions were correct. This is mostly like due to the downward spikes we saw in the number of turns, the number of short/middle turns, and overlap initiation we saw in Figure 1, as well as the upward spike we saw in this level for topic initiation. These distinctive changes may have helped the model predict this level the most accurately in terms of precision. Moving on to recall, the model was most accurate in predicting B1 test-takers (i.e., 0.68). This means that of all the B1 test takers, 68% were accurately predicted to be in level B1. Again, looking at Figure 1, we can see a significant difference between B1 test-takers and all others in terms of the Function 1 features (i.e., interactive turn and initiation). This may have helped the model distinguish B1 test-takers from the rest and therefore, make better predictions as to who should be in this level.

6. Conclusion

The current findings provided evidence to enhance our understanding of the complex nature of interactional competence in the context of the high-stakes speaking assessment. In general, the eleven interaction features were moderately able to distinguish test-takers' CEFR levels with over 50-60 % accuracy. Even though none of the futures showed a straight learner pattern, they were clearly associated with overall interaction scores. In addition, interaction patterns found in the study help strengthen the theoretical basis of interactional competence by empirically investigating the particular interaction features used in candidate's actual interaction.

Some direct implications can be drawn from these findings. ESL instructors can help increase learners' awareness of the interaction patterns (e.g., Function 1 - *Interactive Turn and Initiation)* so that more collaborative conversations can be expected in L2 speaking or test preparation classrooms. Teachers can direct learners' attention to certain interaction patterns explicitly (e.g., turn taking) and show connections between such features and speaking scores. They can further train learners to engage and participate in conversation more proactively. Given that an interaction has multiple communication functions such as argument, discussion, and explanation, identifying and practicing varying communication functions can serve as a good exercise to achieve effective communication goals.

There are limitations to the current study that can be addressed in future research. First, an alternative statistical method can be used to distinguish interaction performance at different levels. We used the canonical discriminant analysis (CDA), but this can be only one way of examining the patterns and prediction of features on outcome variables. Next, future studies can explore relationship between interactive features and other linguistic features, i.e., the task differences. Although studies have often examined the relationship between linguistic features in oral performances and proficiency levels (Brown, Iwashita and McNamara 2005), further research is needed on how different task types affect linguistic features in speaking assessment. Also, more interaction features such as a breakdown of discourse markers and repair strategies can be included with a larger sample size. Addressing these limitations in future research may aid in developing a more robust model for predicting proficiency level based on interaction features.

References

[1] Kang, O., Rubin, D, & Lindemann, S. (2015). Using contact theory to improve US undergraduates' attitudes toward international teaching assistants. TESOL Quarterly, 49, 681-706.

[2] Hall, J. K., & Pekarek, D. S. (2011). L2 interactional competence and development. In J.K. Hall, J. Hellermann & S. Pekarek Doehler (Eds.). L2 interactional competence and development (pp. 1-18). Bristol, UK: Multilingual Matters.

[3] 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.

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

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

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

[7] 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.

[8] Oksaar, E. (1990). Language contact and culture contact: Towards an integrative approach in second language acquisition research. In H. Dechert (Ed.). Current trends in European second language acquisition research (pp. 230-243). Clevedon: Multilingual Matters.

[9] Kasper, G. (2006). Beyond repair: Conversation Analysis as an approach to SLA. AILA (Association Internationale de Linguistique Appliquée) Review, 19, 83-99.

[10] Storch, N. (2001). An investigation into the nature of pair work in an ESL classroom and its effect on grammatical development. Unpublished Dissertation, University of Melbourne, Melbourne.

[11] Storch, N. (2002). Patterns of interaction in ESL pair work. Language Learning, 52 (1), 119-158.

[12] Dimitrova-Galaczi, E. (2004). Peer-peer interaction in a paired speaking test: the case of the First Certificate in English (Unpublished PhD dissertation). Teachers College, Columbia University.

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

[14] May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. Language Assessment Quarterly, 8(2), 127-145

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

[16] Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. Issues in Applied linguistics, 6(2), 5-35.

[17] Schegloff, E., A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D., Tannen (Ed.) Analyzing discourse: Text and talk (pp.71-93). Washington, D.C.: Georgetown University Press.

[18] Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. Applied Linguistics, 11, 189-199.

[19] Gnisci. A. & Bakeman. R.(2007). Sequential Accommodation of Turn Taking and Turn Length: A Study of Courtroom Interaction. Journal of Language and Social Psychology, 26 (3), 234-259.

[20] Nakatani, Y. (2010). Identifying strategies that facilitate EFL Examinees' oral communication: A classroom study using multiple data collection procedures. The Modern Language Journal, 94. 116-136.

[21] Plonsky, L. (2015). Advancing quantitative methods in second language research. Routledge.

[22] Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics. Allyn & Bacon/Pearson Education.

[23] Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks. (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service.