English Teaching, Vol. 78, No. 1, Spring 2023, pp. 63-82
DOI: https://doi.org/10.15858/engtea.78.1.202303.63
http://journal.kate.or.kr

English Oral Proficiency Measured by Holistic and Analytic Assessments in Dialogic and Monologic Tasks*

HyunSook Ko**

Ko, HyunSook. (2023). English oral proficiency measured by holistic and analytic assessments in dialogic and monologic tasks. *English Teaching*, *78*(1), 63-82.

This study compared Korean adult learners' English speech in dialogic and monologic tasks, measuring 57 college students' oral proficiency in terms of holistic and analytic grades. The statistical analyses were focused on two questions: 1) how much do the holistic grades of oral proficiency and analytic grades of the four linguistic features function differently across the tasks? and 2) which linguistic feature has the strongest correlation with the holistic grades of English oral proficiency? The study found significant differences between the two tasks in five aspects (English oral proficiency, topic development, fluency, range, and accuracy). It also showed that a monologic task resulted in lower achievements in those five aspects. In the comparison of the correlations with English proficiency, the range reached the top in both tasks, followed by topic development. The study provides some helpful information regarding which types of speech tasks and linguistic features of EFL oral proficiency are more desirable for targeting in teaching and assessment.

Key words: task types of English speaking, dialogic tasks, monologic tasks, topic development, range

* This study extended the work of Ko (2022) through human analytic scoring and new findings.

** Author: HyunSook Ko, Professor, Department of English Language and Literature, University of Ulsan, #14-514, 93 Taehak-ro, Nam-gu, Ulsan 44610, Korea; Email: hyunsko@ulsan.ac.kr

Received 31 December 2022; Reviewed 15 January 2023; Accepted 17 March 2023

1. INTRODUCTION

Many studies on second or foreign language (L2 as a cover term) education have discussed speaking as one of the core communication skills to master (Fulcher, 2003; Kim, 2015; Ogawa, 2022). Moreover, it is one of the key issues to identify critical features of speaking proficiency in L2. For example, Granena (2019), Housen and Kuiken (2009), and Skehan (1989) claimed complexity, accuracy, and fluency were the main indicators or principal dimensions of L2 proficiency.

However, previous research on speech proficiency has been limited in providing practical suggestions for classroom pedagogy, specifically for promoting L2 learners' achievement on different tasks. The gap includes a lack of information on how similar or different results are produced on a specific task compared to results from other types of tasks as well as which features of L2 speech have a stronger correlation with speech proficiency scores.

Interestingly, Ferrari (2012), Michel, Kuiken, and Vedder (2007), Michel (2011), Mostafa (2021), and Tavakoli (2016) showed that L2 learners tend to speak more fluently in dialogic tasks than in monologic tasks. Moreover, Mostafa's (2021) study showed that highly proficient speakers of English as a second language (ESL) produced more syntactically and lexically complex language in monologic speech, but faster speech in dialogic speech. Such findings of previous research indicate that oral proficiency in English as a foreign language (EFL) should be understood from an integrative perspective to carefully consider the complicated or multi-dimensional relationships between assessment features and task types.

In this context, the current study compares EFL learners' oral proficiency achievements in monologic and dialogic tasks. It investigates the relationships between EFL learners' oral proficiency achievements and linguistic features, such as topic development, fluency, range, and accuracy. Each of these features has been found to be a critical component of L2 oral proficiency in the majority of previous research (Brown, Iwashita, & McNamara, 2005). By offering more pedagogically integrated and specific research, the current study focuses on comparing both the holistic scores of L2 oral proficiency and the analytic scores of linguistic features to demonstrate whether significant differences exist between dialogic and monologic tasks in terms of holistic oral proficiency scores and linguistic features. The current study also discusses which linguistic features have a significant and the strongest correlation with holistic proficiency scores. Ultimately, the results provide practical and helpful information regarding which linguistic features are the priority in when quickly assessing L2 learners' oral proficiency, depending on the type of tasks (e.g., monologic and dialogic tasks).

Specifically, the current research will answer the following three questions:

1) Do Korean adult learners show significant differences between dialogic and monologic tasks in terms of holistic scores of English oral proficiency?
2) Which linguistic features (i.e., topic development, fluency, range, and accuracy) have significant differences between dialogic and monologic tasks?
3) Which linguistic feature has a significant and the strongest correlation with holistic scores of English oral proficiency in dialogic and monologic tasks?

2. BACKGROUND

2.1. L2 Oral Proficiency and Its Assessment Features

Lin (2022) and Pallotti (2009) provided insightful discussion on the definition of L2 oral proficiency. Lin (2022) defined it as "L2 learners' ability to speak their second language to ensure communicative objectives in real-life settings" (p. 1). Interestingly, Pallotti (2009) defined communicative adequacy as "the degree to which a learner's performance is more or less successful in achieving a task's goals efficiently" (p. 596), and some of research on L2 speech has adopted the term in discussing the assessment of L2 speech. Thus, the current study uses L2 oral proficiency as a cover term referring to a communicative competence for L2 speech, including communicative adequacy from Pallotti (2009).

As for assessment features in EFL speech, the current research adopts the those commonly used from Luoma (2004), Roca-Varela and Palacios (2013), and Ulker (2017). These authors argued that there are four critical dimensions successfully characterizing proficiency levels across a variety of English speech tests: topic development (or task completeness), fluency, range, and accuracy. These four features were adopted from the five dimensions discussed in the Common European Framework of Reference for Language (CEFR) by the Council of Europe (2001). Note that, different from CEFR, the rubric does not include interaction (mainly referring to the ability to mainly control turns) as the presence of interaction in monologic speech is not commonly taught in the national curriculum in Korea (Kim, Jung, & Tracy-Ventura, 2017; Park, 2018, 2021). In addition, the scoring rubric of the current study renamed coherence in the CEFR as topic development. Topic development (or task fulfillment/completion) is a more general term to evaluate coherent organization and focus, sociolinguistic appropriateness, and task completeness (Hassanali, Yoon, & Chen, 2015; Koizumi, In'nami, & Fukazawa, 2020; Park, 2012; Roca-Varela & Palacios, 2013; Supakorn, 2017). The four key features of English oral proficiency used in this study were defined as follows:

1) Topic development refers to the ability to build a logical discourse with connectors, other cohesive and coherent devices, and contents sufficiently easy to understand.
2) Fluency refers to the ability to speak at a normal speed or a natural flow without notable hesitations.
3) Range refers to the variety of lexical repertoire, sentence patterns and formulaic expressions used in a speech.
4) Accuracy refers to the ability to make correct use of the language, including intelligible pronunciation and the correct use of grammatical structures.

To summarize previous research on the four features in L2 oral proficiency, Park (2012) reported the highest achievement of task completion (topic development in the current study) when engaging in picture-cued storytelling, explaining about one's favorite TV show, and retelling a story in order. The research also showed the strongest correlation of discourse competence, followed by task completion, with holistic scores of L2 oral proficiency.

Related to fluency, Crossley and McNamara (2013) and Tonkyn (2012) showed that word type counts predicted 61% of the variance in L2 proficiency, while Huensch and Tracy-Ventura (2017) found speech rate and length of pause to be good predictors of L2 proficiency. On the other hand, Vercellotti (2019) reported that, apart from such linguistic measurements of fluency, diversity of clause types (range in the current research) had a significant linear correlation with ESL adult learners' oral proficiency. Meanwhile, Iwashita, Brown, McNamara, and O'Hagan (2008), Révész, Ekiert, and Torgersen (2016), and Tonkyn (2012) found that the use of some grammatical structures (such as subordinate/conjoined clauses, subject verb agreement, tense aspect forms, and primary auxiliaries) had a significant relationship with learners' higher proficiency.

Many results from previous research have presented complicated findings rather than converging to generalized conclusions across task types, measurements of proficiency features, language background, and participants' different ages and levels of proficiency (Bulantová, 2020; Kellogg & Han, 2001; Park, 2012). Thus, the current study focuses on investigating the relationships of holistic and analytic scores across task types for Korean EFL adult learners.

2.2. Holistic and Analytic Assessments

The current study employs both holistic and analytic assessments to compare English oral proficiency between dialogic and monologic tasks. As discussed in Brookhart (2013), Brown (2004), Koizumi et al. (2020) and Park (2012), holistic scoring assigns a single score to elicited responses to given tasks, but its scales focus on the overall task and the discourse ability needed to accomplish the goals of the tasks. Previous research, including Brown (2012), has also emphasized that a holistic rubric is easier and more efficient to use than an analytic one.

On the other hand, Koizumi et al. (2020) pointed out that ESL teachers should carefully select measurement components in an analytic approach, because of the gaps between experimental contexts and classroom contexts in terms of the time and task limits. According to Koizumi et al. (2020), "interactive communication (to the extent of how to actively participate in communication)" and "fluency" in analytic scoring generate statistically stronger correlations with holistic scores. Some specific features of oral proficiency in analytic scoring are closely related to holistic scores, whereas others are not. For example, Iwashita et al. (2008) reported that higher speech rate, higher verb phrase ratio, longer utterance per unit, wider range of word types, and grammatical accuracy successfully predicted EFL oral proficiency among Japanese learners. Meanwhile, Tonkyn (2012) found that speech rate, fluent runs, the length of AS-unit[1], use of primary auxiliaries (i.e., "*do,*" "*be*," and "*have*"), subordinate clauses, the number of word types, and accurate verb phrases were positively correlated with ESL oral proficiency. In addition, Révész et al. (2016) added the number of errors per 100 words as a good predictor of L2 oral proficiency.

Considering these different findings in the previous research, it is necessary to examine whether scores of English oral proficiency functions independently from or similarly to each other, depending on types of tasks and scoring. Thus, the current research compares the results between holistic and analytic scoring. For classroom contexts or the situations of placement tests, particularly when simple and quick assessments are necessary, specific information on the relationship between analytic features and holistic scores may help teachers and test administrators temporarily diagnose learners' level of proficiency in a timely manner. In other words, the results of the current research may provide practical and useful information for ESL teachers who need efficient measurements of learners' speaking achievements in classroom settings. In addition, the study may also contribute with more detailed and helpful information for test developers, particularly on which type of tasks would be more appropriate to use for determining the level of L2 learners' proficiency, depending on their purpose or target features to test.

[1] AS-units refers to sentence level utterances with at least an independent clause with a finite verb and its dependent clauses (Foster, Tonkyn, & Wigglesworth., 2000; Vercellotti, 2019).

2.3. Dialogic and Monologic Tasks

Following Skehan (2001) and McCarthy (2010), the current study defines a dialogic task as a speech task in which a participant is required to interact with a partner or interlocutor, whereas a monologic task is a speech task in which a participant individually delivers a narrative without any interaction with a partner or interlocutor. Many comparative studies on linguistic features in L2 speech have considered these two types of tasks (Ferrari, 2012; Michel, 2011; Michel et al., 2007; Tavakoli, 2016). However, few studies have investigated the relationships between linguistic features and oral proficiency across the task types. Among such research, particularly, Mostafa (2021) and Ko (2022) are particularly worthy of note. Mostafa (2021) examined the speech of 60 ESL speakers, including a majority of Chinese and Arabic speakers, and found a higher articulation rate predictive of higher ESL oral proficiency for a dialogic task as well as a longer mean length of clause and a higher lexical sophistication predictive for a monologic task. Interestingly, the researcher also reported that the phonation time ratio and false starts per 100 words were predictive of ESL oral proficiency for both types of tasks. On the other hand, Ko (2022) investigated Korean adults' English speech and found that word type counts and the number of error-free verb phrases have strong correlations with English proficiency for a dialogic task; for a monologic task, the number of word tokens per minute and the number of error-free verb phrases had strong correlations. Yet these previous studies are limited in providing insightful information on the relationships between L2 oral proficiency and its component features in two common types of scoring, holistic and analytic scoring.

In this context, the current study compares the results of both dialogic and monologic tasks to investigate how linguistic features in analytic scoring relate to L2 oral proficiency measured through holistic scoring. The findings of the current research will provide useful information to enhance the understanding of the relationships of L2 oral proficiency and linguistic features, such as topic development, fluency, range, and accuracy across the tasks.

3. METHODOLOGY

3.1. Participants

Participants were recruited at a university in Ulsan (with a population of about 1,000,000), Korea, through both oral and written introductions presented by the researcher in two courses: "English Reading and Writing," and "English Grammar." After signing consent forms indicating their willingness to volunteer, 57 students in total completed their speech performance on two tasks and received 10,000 KRW (about $7.10 USD) in compensation. The participants (i.e., 55 freshmen and two sophomores) were majoring in English language and literature. Their age averaged 19.04 years and ranged from 18 to 22. Sixteen participants were male and the rest were female. Each participant made an appointment with the researcher and recorded their two speeches in the researcher's office: one for a dialogue completion of a job interview and the other for a monologue of giving opinions upon payment. Although six of the participants did not have any standardized test scores for English, the remaining 51 participants reported averages of 293.73 for reading and 324.51 for listening on the TOEIC. These scores suggest that the participants were intermediate English learners (Boulton, 2008; Chapman, 2006).

3.2. Research Instruments and Data Collection

The current research adopts two kinds of instruments from previous studies. To elicit Korean learners' English speech, the current study followed almost the same formats and procedures as a dialogic and a monologic task from Brown (2004, p. 150) and Park (2012, p. 189), respectively (see Appendices A and B). Both tasks have also been used in Ko (2022), who investigated the functions of word type counts, word token ratio, and verb phrase counts in L2 speech.

Minor modifications were made to improve participants' familiarity with topics or situations. Thus, instead of the conversation between a clerk and a customer in Brown (2004), the dialogic task in the current study presented a job interview asking about personal information. On the other hand, the monologic task asked participants to recount their preferred payment methods in the same way as in Park (2012). Thus, for the dialogic task, a role play of a job interview with blanks were presented, and the participants were requested to make a speech as a role player, filling in the blanks with their answers to the interlocutor's questions. The whole text given as a dialogue in the task was originally a transcript from an authentic job interview between two native English speakers. For the dialogic task in the current study, only the interviewee's comments in the transcript were replaced with the blanks. On the other hand, the monologic task asked participants to designate their preferred payment method (i.e., credit cards or cash) and explain the reasons.

When participants visited the researcher in her office at their appointment time, they were first introduced to the research procedures, which took about five minutes, including small talk to help participants feel comfortable. Then, the dialogic and monologic tasks were presented in order with a one-page written text as well as an oral narration by the researcher. All 57 participants completed their entire speech in each of the two tasks in about three to five minutes, or eight to ten minutes in total. Their speech data were recorded using an MP3 recorder, and the researcher checked the quality of sounds recorded. After no problem was found, each of the audio files of the speech were assigned with an identification number for scoring. The speech files were then rated by one native English speaker with a master's degree and the researcher, who has a doctor's degree. Both raters have more than ten years of English teaching experience at universities in Korea after obtaining their final degree for English teaching in the U.S.

The format of the rubric (see Appendix C) and rating procedures were mainly followed Lee (1995) and Ko (2022), including the two steps for scoring—namely, first compare a pair of speech performances by different speakers to classify each into three main levels (i.e., advanced, intermediate, or beginner) and then refine scores into one of three sub-levels (i.e., plus, neutral, or minus) within a level. As Lee (1995) and Ko (2022) emphasized, such scoring procedures offer the raters some advantages, including reducing their burden of memorizing too many descriptors. The raters used a checkmark (√) next to one of the three options to complete their scoring, which might have helped the raters keep their insights consistent and refined (see Appendices D and E). Finally, each speech sample was judged to belong to one of nine levels.

The same procedures were applied for both the holistic and analytic scoring in order, as follows: Before starting the main scoring, the native English speaker rater completed a training session using the researcher's guide. The two raters (one native English speaker and the researcher) individually read the draft of a rubric and had several online discussions with each other to enhance their clear understanding of it. The raters then independently scored two sample files of the speech data for each session and checked the scoring results with each other. Both raters participated in an online conference to reach a consensus on the scores. The discussion was focused on such scores that were three higher or lower than the other rater's, intending to improve inter-rater reliability in the main study. After confirming that both raters shared the similar calibration and criteria, such as all the scores for the sample data being less than a two-step difference between raters, another two sample files were scored as the second session of scoring. Thus, the final version of the rubric (see Appendix C) was completed through the two raters' cooperation and shared with each other as guidelines. The final version of the rubric was provided from a facilitating perspective rather than as exhaustive lists with a controlling intention. Finally, the main scoring was carried out independently, and all scores were collected; the averages of the two raters' scores were analyzed using SPSS 28.0. The research adopted two approaches to keep the holistic and analytic scoring independent: 1) assigning different identification numbers to speech files across the holistic and analytic scoring, and 2) starting the analytic scoring after completing the holistic scoring of all the participants and the speech files.

3.3. Research Analysis

Using the Kolmogorov Smirnov test, the normality of the holistic scores for the participants' speech was tested. The two raters' scores were found to have significantly deviated from a normal distribution at the 95% confidence level ($p < .01$ for both tasks and for both raters). Thus, to check the inter-rater reliability, the scores were analyzed using the Spearman's rho test. The results showed that all scores had high reliability ($r_s = 0.781$ for the dialogic task; $r_s = 0.817$ for the monologic task).

To determine whether significant differences occurred in English oral proficiency between the two tasks, the Wilcoxon signed-rank tests were conducted on the holistic scores. Likewise, a pair of analytic scores for the same linguistic features across the tasks were examined to determine whether significant differences occurred in the analytic scores between the two tasks. For example, the analytic scores for topic development in the two tasks were tested using the Wilcoxon signed-rank tests.

Finally, correlation analyses were performed for the raters' holistic scores and analytic scores using the Spearman's rho test to determine which linguistic features had a significant correlation and the strongest correlation with holistic scores in each of the dialogic and monologic tasks.

4. RESULTS

4.1. Descriptive Statistics

Table 1 shows the results of the descriptive statistics of the English oral proficiency scores and the linguistic features for the dialogic task and the monologic task.

TABLE 1

Descriptive Statistics of English Oral Proficiency and Linguistic Features (*n* = 57)

	Dialogic Task			Monologic Task		
	Mean	Std. Deviation	Median	Mean	Std. Deviation	Median
Oral proficiency	**3.85**	1.43	**3.50**	3.38	**1.59**	3.00
Topic development	2.81	1.38	2.50	**3.49**	**1.65**	**3.00**
Fluency	2.90	1.50	2.50	**3.52**	**1.90**	**3.00**
Range	2.36	1.42	2.00	**2.90**	**1.58**	**2.50**
Accuracy	2.47	1.27	2.00	**3.35**	**1.63**	**3.00**

Note. Bolded scores represent the bigger one in the comparison of the two tasks.

As Table 1 indicates, the dialogic task shows higher means and medians for oral proficiency in holistic scoring. However, in the analytic scoring, the monologic task shows higher means and medians for all four features. To determine whether such impressions are statistically verified, the Wilcoxon signed-rank tests and the Spearman's rho tests were used.

4.2. Differences in English Oral Proficiency between Two Tasks

To answer the first research questions−namely, whether significant differences exist in English oral proficiency between a monologic and a dialogic task—the Wilcoxon signed-rank test was applied. Table 2 shows the test results.

TABLE 2

Comparison of Oral Proficiency Between Two Tasks (*n* = 57)

	(Scores of Monologic Task)-(Scores of Dialogic Task)				
	Negative Ranks	Positive Ranks	Ties	z	Asymp. Sig. (2-tailed)
Oral proficiency	34	15	8	-2. 264	.024*

Note. (*) represents a significance difference at *p*=.05 level.

As Table 2 indicates, the scores of English oral proficiency scores were found significantly higher in the dialogic task than in the monologic task. It indicates that the monologic task was significantly more challenging for the Korean learners of English and that the Korean learners were judged with lower levels of oral proficiency for the monologic task. Then, to identify which linguistic features of English speech contribute to such a difference, the score differences of four linguistic features between the tasks were examined.

4.3. Differences in Linguistic Features between Two Tasks

To answer the second research questions−namely, whether significant differences in achievements of linguistic features occur between a monologic and a dialogic task—the Wilcoxon signed-rank test was conducted. Table 3 summarizes the test results.

TABLE 3

Comparison of Four Linguistic Features between Two Tasks (*n* = 57)

	(Scores of Task B)- (Scores of Task A)				
	Negative Ranks	Positive Ranks	Ties	z	Asymp. Sig. (2-tailed)
Topic development	17	34	6	-3.242	.001**
Fluency	16	29	12	-2.554	.011*
Range	16	31	10	-2.423	.015*
Accuracy	16	35	6	-3.507	.000**

Note. (*) and (**) represent a significance difference at *p* = .05 level and *p* = .01 level, respectively.

As seen in Table 3, significant differences were found across all four linguistic features (i.e., topic development, fluency, range, and accuracy). Considering these results with the significantly higher scores in oral proficiency on the dialogic task shown in Table 2, these results imply that the participants made more successful speeches during the dialogic task than the monologic task in terms of the four features. Based on the results, the current study recommends monologic tasks, particularly when an assessment mainly aims to finely identify a level of English oral proficiency beyond the beginning level. Monologic tasks may be more useful than dialogic tasks for measuring the four features. On the other hand, when an assessment mainly aims to encourage or motivate learners around the beginning level of proficiency with successful experience, dialogic tasks may be more useful.

Finally, the results summarized in Table 3 verify the participants' lower achievements on the monologic task in terms of the four linguistic features (i.e., topic development, fluency, range, and accuracy) as well as oral proficiency when holistically measured. Such findings seemingly conflict with the higher means of the monologic task for all four features in the descriptive statistics of Table 1. The results in Table 1 present the descriptive characteristics of scores distinguished by two task groups, but they do not explain any trends in matched pairs of scores from the same participants. Meanwhile, the results in Table 3 reflect the statistical differences in a pair of scores from the two tasks by each participant. Thus, the higher achievements on a dialogic task indicated in Table 3 are still valid for comparisons with learners' achievements on a monologic task. Integrating the results summarized in Table 2 and 3, the research indicated that Korean learners produced lower achievements on the monolinguistic task than the dialogic task in terms of topic development, fluency, range, and accuracy. The higher means with the monologic task in Table 1 might be due to its larger individual differences among the participants, compared to the dialogic task. The standard deviations were larger on the monologic task than with the dialogic task (see Table 1).

4.4. Correlations of Linguistic Features with English Oral Proficiency on Two Tasks

To answer the final research question—namely, which linguistic features in the analytic scoring have a significant and strong correlation with holistic scores of English oral proficiency for the two tasks—Spearman's rho tests were conducted. Table 4 shows the results.

TABLE 4

Correlations of Linguistic Features with English Oral Proficiency (***n*** = 57)

Spearman's rho	Dialogic Task		Monologic Task	
	Correlation Coefficient	Sig.	Correlation Coefficient	Sig.
Topic development	.665	.000**	.667	.000**
Fluency	.605	.000**	.599	.000**
Range	.683	.000**	.689	.000**
Accuracy	.625	.000**	.628	.000**

Note. (**) represents significance at the 0.01 level (2-tailed).

As seen in Table 4, all the linguistic features have significant correlations with English oral proficiency in both tasks, with range having the strongest correlation (r_s=.689 for the dialogic task and r_s=.683 for the monologic task), followed by topic development. Such results indicate that range and topic development, in that order, should be prioritized as a target feature to measure Korean adult learners' English oral proficiency across dialogic and monologic tasks. In addition, based on the results of the research, all four features showed a statistically strong correlation with English oral proficiency, following a general discussion on correlation coefficients (Shin, 2014). These findings indicate that all four features may serve as promising predictors of English oral proficiency on each task when assessing Korean adult learners' speeches.

5. DISCUSSION AND CONCLUSION

Based on the Wilcoxon signed-rank test, the current study found that Korean learners achieved higher English oral proficiency scores on a dialogic task than a monologic task. The research also found that the learners achieved significantly higher scores on the dialogic task than the monologic task, particularly in terms of four linguistic features of speech: topic development, fluency, range, and accuracy. These findings confirm that a dialogic speech task in English may be less challenging for Korean adult learners.

Interestingly, such task-dependent significant differences of the current study partially support the works of Ko (2022), Michel et al. (2007), and Michel (2011). Whereas Michel et al. (2007) and Michel (2011) showed higher achievements of fluency and accuracy on dialogic tasks in terms of physical characteristics of speech, including number of errors and speech rate (or ratio of syllables per minute), the current study confirmed such superior achievements in a dialogic task in terms of grades given by human insights of those speech features. Moreover, Ko (2022) reported no significant differences between a dialogic task and a monologic task in terms of some physical measurements representing accuracy and range (e.g., number of error-free verb phrases and mean length per finite clause); meanwhile, the current study found significant differences in terms of human analytic scores in accuracy and range. These complicated findings indicate that further studies need to investigate the relationships between linguistic features and human assessment of oral proficiency with more various tasks and in wider L2 contexts. For example, considering the different results between the previous studies and the current study, further studies may focus on answering a question, whether such conflicting results in accuracy and range are due to insensitivity of physical measurements of speech or to might-be lower reliability of human scoring. Such additional research might help develop more convincing models and procedures of ESL or EFL speech assessment might be developed.

In addition, the results of the current research showed different aspects of EFL learners' proficiency than Mostafa (2021), who demonstrated that advanced ESL speakers produced more complex language on a monologic speech task. The lower achievements with a monologic task in the current study might be due to participants' different levels of English proficiency or the use of different measurements from Mostafa (2021). The participants scored 3.61 out of 9 scales on average on the tasks in the current study, while those in Mostafa (2021) reported 2.24 of 4 on the TOEFL iBT. Another possible explanation for the different results might be the participants' different mother languages in each study (Korean in the current study versus many different languages in Mostafa, 2021). Finally, some exact corresponding features of speech are lacking between the current studies and Mostafa's (2021) study. The current study discussed range, measuring the variety of both lexical repertoire and grammatical patterns. On the other hand, Mostafa (2021) examined complexity, measuring grammatical diversity. Hopefully, future studies can expand the participants' pool and perform an integrative approach to range and complexity, which may provide specific information on the relationships between participants' proficiency levels and their speech features.

Based on the Spearman's rho test, the current study found significant correlations with the holistic scores of EFL oral proficiency for the four linguistic features (i.e., topic development, fluency, range, and accuracy). The emphasis should be on the findings that those four features were determined with both significant and strong correlations with English oral proficiency. Thus, the current study contributes to providing valuable information needed to develop more convincing models of EFL speech assessment through the procedures more directly and faithfully reflecting human raters' insights into holistic oral proficiency. In addition, the current study emphasized topic development in particular, showing the second strongest correlation with oral proficiency. Although rarely discussed in the previous studies, including Ko (2022), Michel et al. (2007), Michel (2011), and Mostafa (2021) with a focus on physical features of speech such as frequency or ratio, the current study in contrast showed that topic development should be considered one of the most promising predictors of EFL oral proficiency in assessment.

Regarding pedagogy, the current study suggests that the type of tasks should be chosen with great caution, depending on the context, including learners' proficiency levels and assessment purposes. Based on the current research, both a dialogic task and a monologic task may similarly function well in terms of topic development, fluency, range, and accuracy when assessing intermediate-level EFL adult learners' speech based on both lexical and grammatical diversity. However, such an approach might not work properly with learners at higher levels, as indicated from the different results in Mostafa (2021). Based on the results, the current study suggests that a dialogic task may be more appropriate for encouraging learners' risk taking at an intermediate level of English learning for Korean adults. A monologic task may also function better for more refined measurements of English oral proficiency for intermediate learners.

However, this study's findings have limits in terms of their generalizability beyond the studied context. In order to determine how contextual factors (e.g., types of tasks, different language background, and learners' proficiency) are involved in English speech assessment, further studies should expand the pool of those factors and integrate the results to suggest more convincing models of EFL or ESL speech development.

Applicable levels: Secondary, tertiary

REFERENCES

Boulton, A. (2008). Looking (for) empirical evidence of data-driven learning at lower levels. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus linguistics, computer tools, and applications: State of the art* (pp. 581-598). New York: Peter Lang.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: ASCD.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English for academic purposes speaking tasks*. Princeton, NJ: Educational Testing Service.

Brown, H. D. (2004). *Language assessment principles and classroom practices*. Harlow, England: Longman.

Brown, J. D. (Ed.). (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. Honolulu, HI: National Foreign Language Resource Center, University of Hawaiʻi at Mānoa.

Bulantová, B. (2020). *Syntactic complexity in the speech of learners of English: Issues in operationalization*. Unpublished master's thesis, University of Karlova, Prague, Czech Republic.

Chapman, M. (2006). An over-reliance on discrete item testing in the Japanese business context. Retrieved on February 13, 2023, from http://www.vantage-siam.com/upload/bulats/file/file-662732081.pdf

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*.Cambridge, England: Cambridge University Press.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning and Technology, 17*(2), 171-192.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277-297). Amsterdam: John Benjamins Publishing Company.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

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

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency: Links between LLAMA and HI-LAB. *Studies in Second Language Acquisition, 41*(2), 313-336.

Hassanali, K., Yoon, S.-Y., & Chen, L. (2015). Automatic scoring of non-native children's spoken language proficiency. In S. Steidl, A. Batiner & O. Jokisch (Eds.), *SLaTE 2015-Workshop on Speech and Language Technology in Education* (pp. 13-18). Leipzig, Germany: ISCA Special Interest Group SLaTE.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473.

Huensch, A., & Tracy-Ventura, N. (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics, 38*(4), 755-785.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Kellogg, D., & Han, Y. (2001). Assessing oral language proficiency: Cognitive traits vs. performance-based categories. *English Teaching*, *56*(3), 45-68.

Kim, H. O. (2015). A historical review of research on speaking in English Teaching. *English Teaching*, *70*(5), 299-328.

Kim, Y., Jung, Y., & Tracy-Ventura, N. (2017). Implementation of a localized task-based course in an EFL context: A study of students' evolving perceptions. *TESOL Quarterly, 51*(3), 632-660.

Ko, H. (2022). Critical linguistic measurements of EFL oral proficiency by Korean adults in dialogic and monologic tasks. *Journal of the Korean English Education Society, 21*(4), 1-25.

Koizumi, R., In'nami, Y., & Fukazawa, M. (2020). Comparison between holistic and analytic rubrics of a paired oral test. *JLTA Journal, 23*, 57-77.

Lee, W.-K. (1995). Assessing Korean university students' spoken English proficiency. *English Testing, 50*(1), 37-63.

Lin, J. (2022). A structural relationship model for L2 oral proficiency, L2 interest, perceived importance of speaking, and out-of-class L2 contact. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688221074027

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

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

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 141-173). Amsterdam: John Benjamins Publishing Company.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259.

Mostafa, T. (2021). *The relationships between second language speakers' oral productions, oral proficiency, and their individual differences: A longitudinal study*. Unpublished doctoral dissertation, Georgia State University, Atlanta.

Ogawa, C. (2022). CAF indices and human ratings of oral performances in an opinion-based monologue task. *Language Testing in Asia, 12*(4), 2-18.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.

Park, T.-J. (2012). *A comparative analysis of the validity of English speaking performance test tasks.* Unpublished doctoral dissertation, Korea University, Seoul.

Park, Y. (2018). Task-in-process during information-gap activities in Korean middle school English. *English Teaching*, *73*(2), 59-86.

Park, Y. (2021). Task type completion in lower level EFL classes: A conversation analytic study. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/1362168820987957

Révész A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828-848.

Roca-Varela, M. L., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey. *International Journal of English Studies, 13*(2), 53-68.

Shin, S.-K. (2014). *Research methods in foreign language education*. Seoul: Hankukmunhwasa.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Skehan, P. (2001). Task and language performance assessment. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167-185). Princeton, NJ: Pearson Education.

Supakorn, S. (2017). *Topic development in Thai EFL classes: A conversation analytic perspective.* Unpublished doctoral dissertation, Newcastle University, England.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 133-150.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners' short-term gains. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 221-245). Amsterdam: John Benjamins Publishing Company.

Ulker, V. (2017). The design and use of speaking assessment rubrics. *Journal of Education and Practice, 8*(32), 135-141.

Vercellotti, M. L. (2019). Finding variation: Assessing the development of syntactic complexity in ESL speech. *International Journal of Applied Linguistics, 29*(2), 233-247.

APPENDIX A

Dialogic Task

Dialogue Completion

Directions: Now read through the dialog for 1 minute. Then, listen and respond.

A: Have you been waiting long?
B: ____________________________________
A: Oh. It’s nice to meet you.
B: ____________________________________
A: When did you graduate from university?
B: ____________________________________
A: What major were you in?
B: ____________________________________
A: Did you enjoy the study?
B: ____________________________________
A: That’s good. Could you tell me about yourself?
B: ____________________________________

APPENDIX B

Monologic Task

Discussions
Directions: Tell me about your opinions about the topic immediately.
Topic: Paying with cash or credit card. Which is better? Explain the reasons why you think so.

(Park, 2012, p.189)

APPENDIX C

Rubric for Holistic Scoring

Categories are defined as follows:

Topic development (or task completeness) implies the ability to build a logical discourse with connectors, other cohesive/coherent devices, and contents sufficiently easy to understand. In other words, it refers to the ability to control the organization and focus, sociolinguistic appropriateness, or task completeness in speech.

Fluency refers to the ability to speak at a normal speed or natural flow without notable hesitations.

Range refers to the variety of lexical repertoire, sentence patterns, and formulaic expressions used in a speech.

Accuracy includes intelligible pronunciation and the correct use of grammatical structures.

APPENDIX C

Rubric for Holistic Scoring (continued)

Level (score)	Key Characteristics
C- (1)	is very difficult to understand
C (2)	is on-topic, but severely lacks elaboration has frequent grammatical errors including incorrect verbs for past tense (**felled*)
C+ (3)	is a little difficult to understand due to limited use of words (**fell* for *dropped*) and grammar, including extra copula *be*
B- (4)	is fully comprehensible, but only when the listener makes some efforts to interpret it
B (5)	is in need of minor elaboration, mainly on topic development
B+ (6)	has few errors
A- (7)	has few repetitions or self-correction
A (8)	shows no hesitation to find words or grammar
A+ (9)	demonstrates full development of the topic

APPENDIX D

Sample of Scoring Sheet (Holistic Scoring)

* Mark with (√) in the blanks below.

	Level	C-	C	C+	B-	B	B+	A-	A	A+
Speaker's ID	(score)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1										
2										
3										
4										
5										
6										
7										
8										
9										

APPENDIX E

Sample of Scoring Sheet (Analytic Scoring)

Number of the First File: __S1__
Number of the Second File: __S2__

* Mark with (√) in the blanks below.

Level	C-	**C**	C+	B-	**B**	B+	A-	**A**	A+
(score)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Topic Development S1									
S2									
Fluency S1									
S2									
Range S1									
S2									
Accuracy S1									
S2									

- modified from Lee (1995)