English Language Teaching; Vol. 15, No. 9; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

The Influence of Partners' L2 Proficiency on Test-Takers' Performance in Paired Oral Assessment

Jiliang Chen[1*], Lanlan Li[1*], Yalin Shen[2*] & Haoxia Li[1]

[1] University of Science and Technology of China, China

[2] Hefei Normal University, China

Correspondence: Jiliang Chen, the Interdisciplinary Research Center for Language Science, University of Science & Technology of China, Jinzhai Road 96, Hefei 230026, Anhui, China.

*These authors contributed equally to this work.

Received: August 3, 2021 Accepted: July 6, 2022 Online Published: August 22, 2022

doi: 10.5539/elt.v15n9p69 URL: https://doi.org/10.5539/elt.v15n9p69

Abstract

This study is intended to explore the influence of partners' L2 proficiency on test-takers' performance and on their interaction patterns in paired oral assessment, and it tries to analyze the underlying causes on the basis of Vygotsky's theory of the zone of proximal development (ZPD) and Storch's patterns of interaction. An experiment including three rounds of tests was carried out with 12 students who were selected out of 60 from a top university in China. The findings showed that the high-level test-takers received the highest scores when paired with middle-level partners; the middle-level test-takers scored the highest and the lowest when paired with high-level partners and low-level partners respectively; the low-level test-takers had basically the same scores when paired with those of the other two different levels. The results could be partly explained by the fact that the higher level partners stimulated the lower level test-takers to the upper limit of ZPD through language interaction, but the mechanism of ZPD might not work when partners' proficiency levels were much higher than the ceiling of test-takers' ZPD. Besides, test-takers' performance was mediated by pair interaction patterns. The implications of the study not only deliver suggestions for fairness of paired oral assessment but also provide alternatives for test-takers to achieve higher scores in a paired speaking test by choosing partners of different L2 proficiency levels.

Keywords: paired oral assessment, partners' L2 proficiency, test-takers' performance, zone of proximal development, patterns of interaction

1. Introduction

It has been a widely accepted premise in second language (L2) assessment that a learner's communicative competence has to be given the chance to be displayed and then evaluated. The increased demand for performance tests has led to the practice of using language proficiency interviews (LPI) in the assessment of oral proficiency. LPI typically involves a face-to-face interview between a test-taker and an interviewer. However, it has also incurred criticism, which is mainly related to the limited range of samples of speaking ability, as it can only elicit a limited domain of interaction. Additionally, the asymmetry of the power relations between the test-taker and the interviewer may distort the language subsequently used in a conventional assessor-assessed arrangement. The need for speaking tests to provide a better basis for oral language sampling with less asymmetry between participants has therefore led to the use of "alternative organizational arrangements" (Skehan, 2001, 2018) in speaking tests and the introduction of paired or grouped test tasks.

In the language learning and teaching context, paired speaking tests are usually preferred and often selected in oral English assessment for the following reasons: first, they serve as a tool for stimulating the development of oral L2 proficiency by motivating students to collaborate more fully during classroom activities (Saville & Hargreaves, 1999); second, oral communication between peers is a feature of many classroom speaking tasks, and the use of paired testing is well suited to educational contexts where the pedagogical focus is fully or partially task-based. (Long & Crooks, 1992). Paired oral assessment has gained increasing popularity as a useful tool for assessing learners' speaking ability (East, 2015) because it elicits authentic verbal interaction and greatly resembles communicative classroom activities (Philp, Adams, & Iwashita, 2014)

A paired oral assessment can be viewed as a performance that is the result of an interaction between test-taker, test, and partner, which is then judged by one or more raters, who apply a rating scale and produce a score (Mcnamara, 1996). Despite the simplicity of this basic concept, its realization in actual assessment can be extremely complex. For example, performance may vary in response to various tasks, different partners, or specific combinations of task, test-taker, and partner, all of which may ultimately influence the final scores. As viewed from the perspective of sociocultural theory, performance was jointly constructed. It was not a solo performance but rather a socially mediated performance with language mediating the interaction. The participants might exercise differential responsibility, but the process and the product is a joint construction (Swain 2000, 2013). The co-constructed nature of interaction in a test situation posed challenges for language test developers in terms of rating and fairness, as test-takers' and partners' performances were dependent upon each other. As Swain (2001a) states, “in a testing situation, who one is paired or grouped with, is not unimportant”.

Some studies found partner effect in the context of paired testing (Iwashita, 1996, 2001; Watanabe & Swain, 2007; Davis, 2009; Son, 2016), but no consensus has been reached so far. Iwashita (1996) found that when Japanese students were paired with high-level peers, the average scores of high-level students and low-level students increased by 13% and 53% respectively. According to Iwashita, the linguistic ability of partners may affect language output and oral performance. Watanabe and Swain (2007) found that the average scores of testees who cooperated with partners with higher language levels were slightly lower than those who cooperated with peers with lower language levels. Davis (2009) investigated the influence of language proficiency on the oral performance of 20 Chinese university freshmen and found that although low-level testees had more output when they cooperated with high-level partners, generally speaking, the language level of the interlocutors had no significant impact on the scores. Son (2016) obtained similar results, however, she found that low-proficiency test-takers produced fewer words when paired with high-proficiency partners.

To date, many researchers have explored the influence of partners’ L2 proficiency levels on testees’ performance in paired oral assessment, but few studies focused on the fluency, amount of production, lexical diversity as well as patterns of interaction in this testing context. This study tries to examine the influence of partners’ L2 proficiency levels on test-takers’ performance in terms of scores, fluency, amount of production, lexical diversity and their patterns of interaction, and explores its underlying causes on the basis of Vygotsky’s zone of proximal development (ZPD) and Storch’s interaction patterns. This study aims to address the following questions:

Question 1: Is partners’ proficiency level associated with test-takers’ scores in paired speaking tests?

Question 2: How do partners’ proficiency levels affect test-takers’ fluency, amount of production and lexical diversity in paired speaking tests?

Question 3: What is the relationship between partners’ proficiency levels and patterns of interaction in paired speaking tests?

2. Theoretical Rationale

2.1 Zone of Proximal Development and Scaffolding

According to Vygotsky (1978), the zone of proximal development (ZPD) is defined as “the distance between the (child’s) actual development level as determined by independent problem solving and the level of potential development as determined through problems solving under adult guidance or in collaboration with more capable peers”. However, since the relationship between a child and an adult (or a teacher) is similar to the relationship between a student and a more capable peer, the concept of ZPD nowadays has been increasingly used in peer-peer interaction.

American educators interpreted ZPD as an important diagnostic means, an indicator of a child’s hidden potential that became visible only in cooperation with expert others (Rogoff, 1984, 1990; Wertsch, 1985; Newman, Griffin, & Cole, 1989; Moll, 1990; Ellis, Skehan, Li, Shintani, & Lambert, 2020). Rogoff (1990) introduced the term “scaffolding” that extended the concept of the ZPD to indicate the implicit, emotionally engaging, evolving, developing contract between a learner and a teacher. To elaborate the metaphoric term, she introduced “guided participation”, indicating the reciprocal risk that the two placed themselves under when the teacher erected scaffolds and the learner mounted the scaffolding. For Rogoff, instruction might be defined as any joint activity that involved a transaction between neophyte and expert for the purpose of constructing shared understanding. Many researchers have applied the concepts of ZPD and scaffolding to second language education and acquisition (Donato, 1994; Ohta, 2000; McCafferty, 2013, 2018; Khaliliaqdam, 2014). They hold the view that scaffolding in language learning, consisting of supportive behaviors provided by the expert or the more

knowledgeable peer, could facilitate the learner's progress to a higher level of language development when the learner attended to form and meaning.

2.2 Patterns of Interaction

Although the role of interaction involving L2 test-takers or learners has been debated for over two decades, much of the research has focused heavily on quantitative accounts of linguistic behavior, with little attention given to the socially constructed nature of interaction. However, the nature of interaction and its significance have started to receive attention. One of the key studies was Storch's (2001b, 2002) longitudinal investigation into the nature of dyadic interaction in an adult ESL classroom. She proposed the 'patterns of interaction' framework which aims to describe the participants' position in a dialogue and explain the effect that it may have on knowledge construction. She classified four distinct patterns of dyadic interaction among pairs: collaborative, dominant/dominant, dominant/passive, and expert/novice. Two indexes were used to distinguish these four patterns: equality and mutuality. Equality described more than merely an equal distribution of turns, but an equal degree of control over the direction of a task. Thus high equality was evident in interactions where both participants took directions from each other. Mutuality referred to the level of engagement with each other's contribution. High mutuality described interactions that were rich in reciprocal feedback and a sharing of ideas. Graphically, these four patterns could be represented by the four quadrants formed by two intersecting axes, as shown in Figure 1.

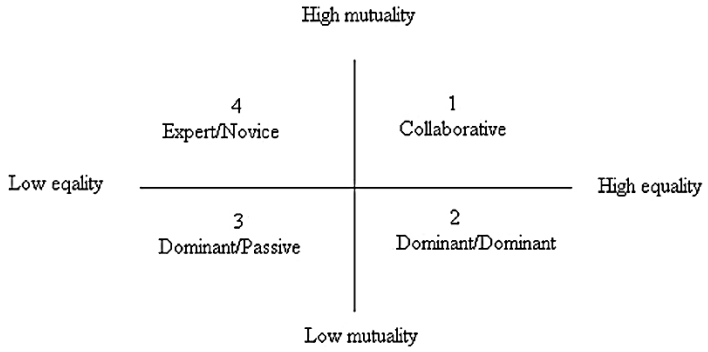

Figure 1. A Model of Dyadic Interaction

In the collaborative pattern, both learners worked together throughout the task with active negotiation and useful assistance, and ended with acceptable resolutions for both participants. The dominant/dominant pairs, however, showed unwillingness or little engagement with each other's contribution, and they were marked by a high level of disagreement and inability to reach a consensus. In the dominant/passive pattern, the dominant participant took an authoritarian stance and controlled the task, while the passive one adopted a more subservient role and made few contributions. For the expert/novice pairs, the expert controlled the task, but actively encouraged the novice to participate in the task. Furthermore, Storch revealed a model of dyadic interaction. The collaborative type had high mutuality and high equality, the dominant/dominant type had low mutuality and high equality, the dominant/passive type had low mutuality and low equality, and the expert/novice type had high mutuality and low equality.

3. Method

3.1 Participants

Twelve out of sixty students in two English classes from a top university in China were selected by their experienced English teacher to represent high, middle, and low oral English proficiency levels, with their current EFL class level, English learning history and former oral English proficiency scores taken into consideration. Then they undertook a pretest monologue task of the shorter version of the College English Test (Band Four) (CET-4), an authoritative national English examination for Chinese college students. The authors intentionally placed pairs of the same gender to eliminate the gender influence, considering Gass and Varonis' (1986) findings that Japanese men seemed to dominate conversations when working with women. The authors also tried to select participants from the same class to minimize the age gap and gave them warm-up time to get familiar with each other. Table 1 demonstrates each student's gender, label, and pretest score.

Table 1. Participants and Pre-test Scores

		High		Middle		Low	
Male	Labels and scores	H1	4.9	M1	4.0	L1	3.2
		H2	5.0	M2	3.7	L2	3.0
	Mean scores	4.9		3.9		3.1	
Female	Labels and scores	H3	4.4	M3	3.5	L3	3.4
		H4	4.1	M4	3.7	L4	2.8
	Mean scores	4.3		3.6		3.1	

3.2 Instruments and Procedures

The monologue and paired speaking test topics were retrieved from the authentic tests of College English Test Band 4 (CET-4). The data were collected immediately after their English class in the language lab room. The timeframe and data collection procedures are shown in Table 2.

Table 2. Timeframe and Data Collection Procedures

Time	*Tasks*
5 minutes	Stage 1: warm-up task
3 minutes	Stage 2: Pre-test monologue task
15 minutes	Stage 3: Paired tasks in the speaking test
5-10 minutes	Stage 4: Post-test interview

Stage 1: Warm-up task. At the beginning of the session, all the upcoming tasks were described and the participants were given a 5-minute self-introduction activity to get familiar with their partners.

Stage2: Pre-test monologue task. The participants were assigned a topic for monologue. They had two minutes to prepare and three minutes to talk.

Stage 3: Paired tasks in the speaking test. The participants then performed three paired tasks with the partners of three different proficiency levels. Table 3 shows the paired test arrangement. Their utterances were tape-recorded and transcribed in CHAT format.

Table 3. Paired Test Arrangement

Male	Topic1	H1-M1	H2-M2	L1-L2
	Topic2	M1-L1	M2-L2	H1-H2
	Topic3	H1-L1	H2-L2	M1-M2
Female	Topic1	H3-M3	H4-M4	L3-L4
	Topic2	M3-L3	M4-L4	H3-H4
	Topic3	H3-L3	H4-L4	M3-M4

Stage 4: Post-test interview. A post-test interview was conducted with each participant after the paired test tasks. This method was used to incorporate each student's behavior during his/her interaction in order to have a better understanding of the nature of interaction which might not be clear from the tape recordings and transcripts alone, to detect characteristics representing a novice, expert, collaborative, dominant, and passive partner which Storch (2002) categorized, and to find out the participants' personal feelings about their interactions.

3.3 Assessment Criteria

The test-takers were scored on a 5-point Likert scale in terms of three task-independent components: code command, cognitive operation and communicative adaptation (Brown, Hudson, Norris, & Bonk, 2002). Code command depends on the kind of language and information that is involved in a successful task performance. Under the concept of code command should be understood the structure of the language relevant to the tasks, including pronunciation, vocabulary, grammar and so forth. Cognitive operation essentially involves the amount and kind of information processing that a testee must engage in for successful performance on the task. It should be understood to involve manipulation of task elements towards the accomplishment of the task, such as organizing or reorganizing information, completion of necessary aspects of tasks and so on. Communicative

adaptation is determined by the type of communicative language activity and by a number of moderator variables that can drastically influence the relative difficulty of a particular communicative act.

3.4 Measures

Both the quality and quantity of the subjects' performance were measured in terms of fluency, amount of production and lexical diversity.

Fluency measures. Fluency was described by "repair fluency", which was the number of repairs in the speeches in the following variables: 1) reformulations; 2) grammar errors; 3) replacements; 4) word repetitions; 5) pauses, including mid-clause pauses like "uh", "um". The number of repairs was the sum of the above variables.

Amount of production measures. The total number of words was measured in "tokens", which was calculated by CLAN software.

Lexical diversity measures. The authors adopted a new measurement of lexical diversity free from the disadvantages of type-token ratio (TTR), that is, a larger number of tokens in a large sample produced lower TTRs than in a small sample (Chotlos, 1944; Richards, 1987). The new measurement was developed and proved by Malvern and Richards (2002). It was a mathematical equation relating TTR to token size in terms of a third variable, a parameter referred to as "D":

$$TTR = \frac{D}{N}\left[\left(\sqrt{1+2\frac{D}{N}}\right)-1\right]$$

In this study, D was calculated by the D-optimum average.

3.5 Interrater Reliability

All test-takers' performances were scored by two experienced English teachers from the University of Sci. and Tech. of China where the data were collected. They assessed three pieces of monologue tapes together to reach a general agreement on scale standards, and then they judged all the other performances. The final scores were the mean of their marks. The interrater reliability is 0.756.

3.6 Categorization of Patterns of Pair Interaction

For the analysis of the transcribed talk for patterns of pair interaction, the authors adopted Storch's (2001b) "patterns of dyadic interaction and associated traits" as a guideline. Storch described and classified the four distinct patterns of interaction in terms of (1) pattern of contribution; (2) decision-making behavior; (3) nature of assistance; and (4) discourse and linguistic features. Based on the coding of the transcripts for each category, the authors attempted to identify each pair as fitting one of the four patterns, and a post-test interview was also used to supplement this analysis.

4. Results

To address the research questions, the results from analyses of variance are presented below.

Question 1: Do partners' proficiency levels affect test-takers' scores in paired tests?

An analysis of variance (ANOVA) was used to examine the relationship between test-takers' scores and their partners' language proficiency. The influence of the partners' L2 proficiency levels on test-takers' average raw scores did not reach statistical significance (high-level test-takers: $F(2,11)=0.184$, $p=0.835$; middle-level test-takers: $F(2,11)=0.646$, $p=0.547$; low-level test-takers: $F(2,11)=0.023$, $p=0.977$). However, judging from the patterns of scores received by the high-, middle- and low-level test-takers in the three paired oral tasks, the test-takers were influenced by their partners from 0.0% to 10.9% (see Figure 2).

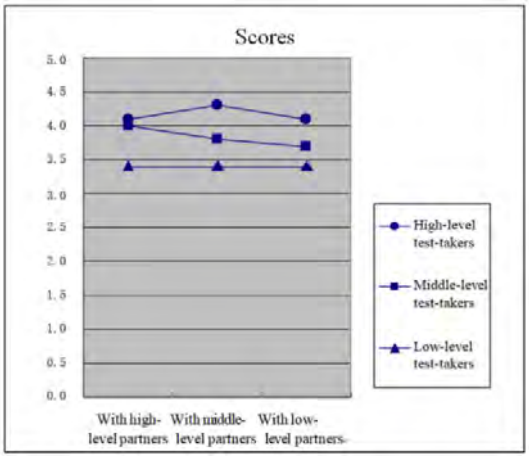

Figure 2. Patterns of Scores Received by the High-, Middle- and Low-level Test-takers in the Three Paired Speaking Tasks

The high-level test-takers received the highest scores when paired with middle-level partners, but did not reach the peak when paired with those of the same level; the middle-level test-takers scored the highest and the lowest when paired with high-level partners and low-level partners respectively while the low-level test-takers had basically the same scores when paired with partners of the other two different levels. The results seemed to contradict Iwashita's (1996) findings that Japanese students with both high and low levels received an increase of 15% and 50% in mean scores when working with higher level partners. The results also seemed to be incompatible with Larry Davis's (2009) findings that there was an increase of 3% in mean scores for high-level test-takers and a decrease of 5% in mean scores for low-level test-takers.

In order to address research question 2, we divided it into the following three sub-questions.

Question 2a: Do partners' proficiency levels affect test-takers' fluency of the language produced?

To answer this question, the number of repairs in the performance of the test-takers was calculated. The statistics show that high-level test-takers produced significantly fewer repairs (reformulations, grammar errors, replacements, repetitions and pause designators) when paired with middle-level partners ($p<0.001$), but middle-level and low-level test-takers did not have such significance (middle-level: $p=0.835$; low-level: $p=0.922$).

As shown in Figure 3, the high-level test-takers had a similar pattern to that in Figure 2. They had the least repairs when paired with middle-level partners, and received the highest scores correspondingly. Middle-level test-takers' pattern was almost similar to that of scores in Fig. 2, except that they had slightly more repairs when paired with high-level partners. Low-level test-takers slightly reduced the number of repairs gradually when paired with the high-, middle-, and low-level partners.

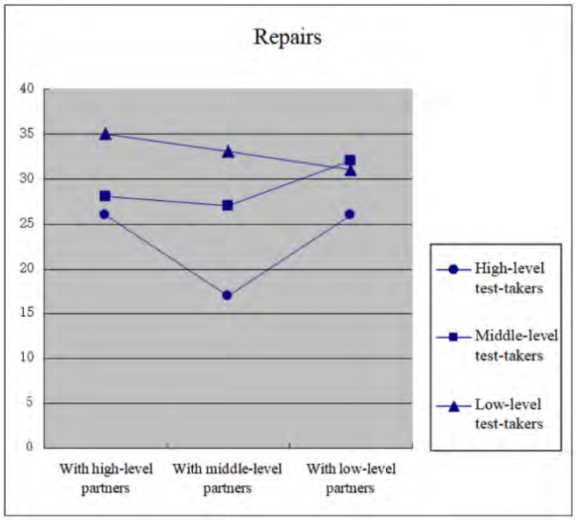

Figure 3. Patterns of Repairs of the High-, Middle- and Low-level Test-takers in the Three Paired Oral Tasks

Question 2b: Do partners' proficiency levels affect test-takers' amount of production?

Generally, partners' proficiency levels had no significant influence on test-takers' amount of speech produced high-level test-takers: $F(2,11)=0.093$, $p=0.912$; middle-level test-takers: $F(2,11)=1.069$, $p=0.383$; low-level test-takers: $F(2,11)=0.068$, $p=0.934$).

However, judging from the number of tokens (see Figure 4), high-level test-takers produced more tokens when paired with middle- and low-level partners, middle-level test-takers' tokens declined sharply when paired with middle-level partners, and low-level test-takers' tokens decreased slightly when paired with low-level partners. This finding was similar to Iwashita's (1996) and Davis' (2009) findings, in which lower-proficiency candidates produced more speech when working with higher-proficiency partners, but higher-proficiency students did not show such an effect.

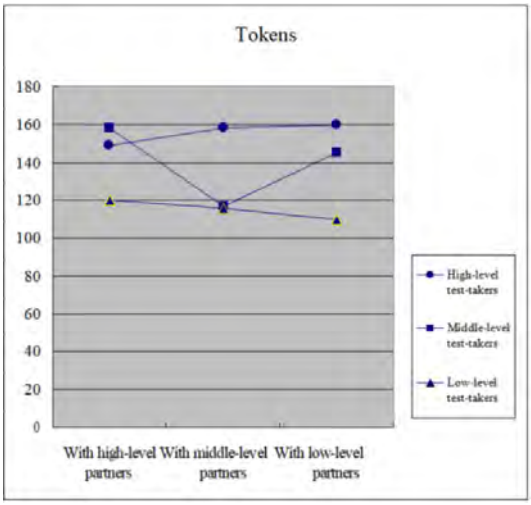

Figure 4. Number of Tokens Produced by the High-, Middle- and Low-level Test-takers in the Three Paired Speaking Tasks

Question 2c: Do partners' proficiency levels affect test-takers' lexical diversity?

Lexical diversity in terms of the D-optimum average generated a similar pattern to that of the amount of production. There was also no statistical significance to prove that partners' proficiency levels had an impact on test-takers' D-optimum average produced (high-level test-takers: $F(2,11)=0.483$, $p=0.659$; middle-level test-takers: $F(2,11)=1.386$, $p=0.299$; low-level test-takers: $F(2,11)=0.065$, $p=0.937$).

The patterns of the D-optimum average revealed that the high-level test-takers produced the highest D-optimum average when paired with high-level partners; middle-level test-takers produced the highest D-optimum average when paired with high-level partners and touched the bottom when paired with middle-level partners; and low-level test-takers' D-optimum average dropped slightly when paired with middle-level partners (see Figure 5).

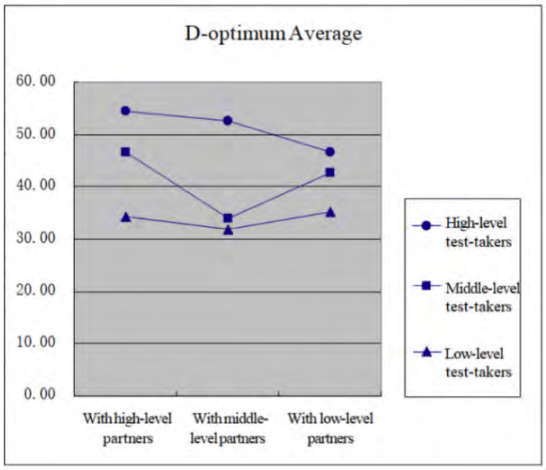

Figure 5. Patterns of the D-optimum Average Produced by the High-, Middle- and Low-level Test-takers in the Three Paired Speaking Tasks

Question 3: Are the partners' proficiency levels associated with the patterns of interaction in paired speaking tests?

When patterns of interaction were examined on the individual level, a number of patterns emerged. As seen from Table 4, the high-level partners produced mainly collaborative interactions (9 of 12 dialogues). In the remaining three pairs, when the high-level partners were interacting with low-level test-takers, two were expert/novice type and one was dominant/passive type. The middle-level partners also produced primarily collaborative interactions (10 of 12 dialogues). These results suggested that high- and middle-level individuals tended to orient to a collaborative interaction pattern if their partners were capable of participating on an equal basis.

A greater variety of patterns of interaction were seen in the low-level group. Collaborative interactions were still the most common (7 of 12 dialogues), but dominant/passive and expert/novice types accounted for 5 out of 12 dialogues.

Table 4. Patterns of Interaction Produced in Paired Tasks for Groups with High, Middle and Low Levels

Gender	Name	Level	Partner	Partners' level	Pretest score	Score	Interaction type
Male	ZSZ	H	LUU	H	4.9	4.7	collaborative
Male	ZSZ	H	XJH	M	4.9	4.7	collaborative
Male	ZSZ	H	LIU	L	4.9	4.4	expert/novice
Male	LUU	H	ZSZ	H	4.9	4.3	collaborative
Male	LUU	H	TAN	M	4.9	5.0	collaborative
Male	LUU	H	LIG	L	4.9	4.6	dominant/passive
Female	WYT	H	WAN	H	4.3	3.9	collaborative
Female	WYT	H	ZNN	M	4.3	4.0	collaborative
Female	WYT	H	XUY	L	4.3	3.5	expert/novice
Female	WAN	H	WYT	H	4.3	3.5	collaborative
Female	WAN	H	ZHA	M	4.3	3.5	collaborative
Female	WAN	H	WEN	L	4.3	3.8	collaborative
Male	XJH	M	ZSZ	H	3.9	3.9	collaborative
Male	XJH	M	TAN	M	3.9	3.5	expert/novice
Male	XJH	M	LIU	L	3.9	3.4	collaborative
Male	TAN	M	LUU	H	3.9	4.4	collaborative
Male	TAN	M	XJH	M	3.9	4.1	expert/novice
Male	TAN	M	LIG	L	3.9	4.0	collaborative
Female	ZHA	M	WAN	H	3.6	3.9	collaborative
Female	ZHA	M	ZNN	M	3.6	3.9	collaborative
Female	ZHA	M	WEN	L	3.6	4.0	collaborative
Female	ZNN	M	WYT	H	3.6	3.6	collaborative
Female	ZNN	M	ZHA	M	3.6	3.6	collaborative
Female	ZNN	M	XUY	L	3.6	3.4	collaborative
Male	LIU	L	ZSZ	H	3.1	4.0	expert/novice
Male	LIU	L	XJH	M	3.1	3.5	collaborative
Male	LIU	L	LIG	L	3.1	3.6	expert/novice
Male	LIG	L	LUU	H	3.1	2.8	dominant/passive
Male	LIG	L	TAN	M	3.1	3.4	collaborative
Male	LIG	L	LIU	L	3.1	2.9	expert/novice
Female	WEN	L	WAN	H	3.1	3.7	collaborative
Female	WEN	L	ZHA	M	3.1	3.4	collaborative
Female	WEN	L	XUY	L	3.1	3.4	collaborative
Female	XUY	L	WYT	H	3.1	3.1	expert/novice
Female	XUY	L	ZNN	M	3.1	3.1	collaborative
Female	XUY	L	WEN	L	3.1	3.5	collaborative

The effect of patterns of interaction on scores was not consistent, for example, the collaborative type did not always generate higher scores for test-takers, and the expert/novice type did not always impose lower scores for low-level test-takers.

5. Discussion

This current study found that partners' proficiency levels generally did not have great effects on the test-takers' performance in the paired speaking test. Although without statistical significance in their average raw scores, the test-takers' scores were influenced by their partners' proficiency to some extent. The increase or decrease in test-takers' scores when they were paired with partners of different L2 proficiency levels could be explained by Vygotsky's theory of ZPD. The more competent partners, whose proficiency level surpassed the actual level of the test-taker, stimulated the development of test-takers' language ability through the medium of language and interaction, which was most evident in the performances of the middle-level group. For the low-level group, the high- and middle-level partners' performances might exceed the ceiling of their ZPD so that they were inaccessible for the low-level group to learn from, therefore they did not generate an obvious rise or fall in the low-level group's scores. Apart from partners' L2 proficiency, test-takers' scores were also influenced by their interaction patterns, which was most evident in the high-level group. When the high-level test takers were paired with the same level partners who were not more competent than the test takers, little development could be stimulated by the partners. When the high-level group was paired with middle-level partners, their interaction types were collaborative, in which learners were willing to offer and engage with each other's ideas, and during the negotiation, alternative views were offered and discussed, often leading to resolutions that seemed acceptable to both participants. Hammond and her colleagues (2002) pointed out the crucial role of language in scaffolding. Through the language in negotiations, the pairs were actively involved and built knowledge together. The knowledge was subsequently appropriated and internalized by members of the pair. Such talk in turn reflected cognitive processes such as noticing the gap, hypothesis formulation, restructuring and understanding which were essential for successful L2 acquisition (Brooks, 2009; Davis, 2009). When the high-level group was interacting with low-level partners, two were expert/novice type and one was dominant/passive type. If there was very little negotiation taking place, such as in the dominant/passive pair, most of the decisions concerning language choices were imposed by the dominant participant, with little input sought from or offered by the passive participant. Thus there would be relatively few instances suggesting internalization and a transfer of knowledge, because there were so few instances where knowledge was co-constructed in ZPD. The overall insignificance of the data could be explained by the limited time of the interaction in the test and insufficient learning time for the test-takers.

According to the statistical results concerning repairs, tokens and D-optimum, the influence of partners' proficiency levels on test-takers' performance in fluency, amount of production and lexical diversity is generally not considerable either, but we also have some mixed findings as mentioned previously, especially with regard to the low-level group's amount of production and lexical diversity, and the high-level group's lexical diversity.

High-level test-takers' pattern of fluency in terms of repairs was found correlated with the score pattern, because they had the minimum amount of repairs when they received the highest scores. Middle-level test-takers did not have the least repairs when paired with high-level partners, partly because they still had many repetitions and self-corrections to complete grammatically correct sentences and were still hard pressed to achieve fluency within a short time of the oral tasks. For example, in the following excerpt from CHAT format, XJH, the middle-level student, was obviously having many repetitions and self-corrections when he was paired with high-level student ZSZ.

*XJH: Oh, yeah, but <if you don't have some> [//] <if you> [/] um if you have good personal relationship, you shouldn't be misunderstood by others.

XJH: If you have a good relationship with your teammates, you [/] you can understand good [: well] [s:r] each other.

*XJH: So sometimes you <shouldn't be> [/] shouldn't be misunderstood.

XJH: And<if you> [/] if you don't have some ability in some ways, you [: your] [s:r-ret] your teammates will help you in the way.

*ZSZ: Yeah, I got it.

*ZSZ: Ok [/] ok [/] ok, I accept that, I accept the opinion of the interpersonal relations is more important, because there are also many examples in China that people may not have the higher ability in the special fields.

Low-level test-takers had slightly fewer repairs when paired with low-level partners, which was partly due to their choice of short sentences. For example:

*LIG: And [/] and I think, um, and I think, uh, if we want (to) succeed, <we should> [//] <we must> [/] we must do hard work and the cooperation is [/] uh, is also a most, imp@pm, uh.*LIU: most important.

*LIG: most important.

*LIU: ability.

*LIG: ability.

*LIU: Yes.

*LIG: Because, uh, many work um, need cooperation.

*LIU: Yes.

*LIG: Such as um [/] such as um.

*LIU: A way we talk with each other [=! laughing].

*LIG: 0[=! laughing].

*LIU: Um, and we talk about it, and I think the most important thing is hard work, because (in) many works we should uh, have our patience.

*LIG: Yeah.

Both LIU and LIG used short sentences to communicate with each other, which naturally reduced repairs.

The amount of speech production in terms of tokens produced by the test-takers also varied when paired with the partners of different proficiency levels. A possible interpretation of this result was that lower level test-takers benefited from working with higher level partners who were more skillful at maintaining a conversation or otherwise eliciting speech from their partners, while higher level students were capable of producing speech without such support.

With an exception, the middle-level test-takers had the lowest amount of production when paired with middle-level partners. This was influenced by the pairs' patterns of interaction. When the middle-level group was paired with high- or low- level partners, their interaction types were collaborative, in which learners could benefit from discussions and negotiations through which the pairs were actively involved and building knowledge together. While paired with middle-level partners, not only their partners were not more capable peers who may stimulate development, but also half of their interaction patterns were expert/novice. In such a pattern, although one participant seemed to take more control over the task, unlike the dominant/passive scenario, this participant acted as an expert who actively encouraged the other participant (novice) to participate in the task. Hence, they took more time to rectify their own or their partners' errors, and supplement partners' production.

As for lexical diversity, the D-optimum average generated a similar pattern to that of the amount of production. Many researchers (Donato, 1994; Lantolf, 2000; Ohta, 2000; Storch, 2017) hold the view that scaffolding, the supportive behaviors provided by the more knowledgeable peer, can facilitate the learner's progress to a higher level of language development. That is, the dialogic process between a teacher or other capable peers and a learner can facilitate the learner to reach the actual development level. Thus, the more competent the partner is, the more skillful he or she will be at eliciting language from the test-taker, which was most evident in the performances of the high-level group.

The middle-level test-takers' lexical diversity was similar to that of language production, both of which were affected by the patterns of interaction. When the middle-level group was paired with high- or low-level partners, their interaction types were collaborative, in which divergent views were discussed and offered in various lexical forms. While paired with middle-level partners, not only their partners were not more capable peers who may stimulate development, but also half of their interaction patterns were expert/novice. Since Vygotskian theory was built upon the Piagetian idea of the child as an active learner, with emphasis on the role of social interaction in learning and development, the quality of learner-teacher or capable peer interaction was seen as crucial when scaffolding learner's learning (Bodrova & Leong, 1996; Fleer, 1992, 1995; Tharp & Gallimore, 1988). While scaffolding can assist children it may also, at times, hinder learners in demonstrating their full range of knowledge (Donovan & Smolkin, 2002). Thus, scaffolding, when understood as direct instruction in the expert/novice pattern, might become counterproductive.

The low-level group reached the highest D-optimum average when paired with low-level partners. This is because scaffolding can also be jointly accomplished by students or novices. Lantolf (2000) and other scholars (Engestrom, 1996; Wells, 1999) have explored peer scaffolding. Their research has demonstrated that learners were able to scaffold each other quietly and effectively through the use of a variety of interactive strategies that appear to be facilitative to ZPD. The implication of their work was that learning could emerge in the absence of a recognized expert. Expertise could also be collaboratively constructed in the talk that occurred among learners who shared the goal of working out a solution to a problem. With respect to language learning, Swain (1995) indicated that dialogues among learners could be as effective as instructional conversations between teachers and learners. Donato (1994, p. 40) pointed out that scaffolding in language learning could be achieved by a language participant "by means of speech, supportive conditions in which the novice could participate, and extended current skills and knowledge to higher levels of competence". Therefore, the supportive conditions did not necessarily have to be created by a teacher or a competent peer and scaffolding could also be jointly accomplished by students or novices.

The present study also found that test-takers' interaction patterns can be affected by their L2 proficiency level. Individuals with higher language proficiency tended to orient to a collaborative interaction pattern in which both learners work together throughout the task completion process and help each other. According to Vygotsky, language was the tool of development, and language used by the test-takers facilitated the co-construction. And in the utilitarian approach's view, participation in collaborative activities was not only mediated by language (as in communication and speech) but also guided by expert others. Therefore, with higher language proficiency, the middle- and the high-level partners are more capable of involving test-takers and themselves equally and mutually into the paired oral task.

6. Conclusion

This study set out to explore the influence of partners' L2 proficiency levels on test-takers' performance and on their interaction patterns in paired oral assessment. Although it did not reach statistical significance, partners' proficiency levels had some influence on test-takers' performance, which was most evident in the middle-level group's scores, low-level group's amount of production and lexical diversity, and high-level group's lexical diversity. In addition, patterns of interaction, which were influenced by the partners' L2 proficiency, in turn, affected test-takers' performance, which was most evident in the high-level group's scores and the middle-level group's production and lexical diversity. The results could be explained by Vygotsky's theory of ZPD and Storch's patterns of interaction. Scaffolding, the supportive behaviors provided by the more knowledgeable peer, could facilitate the learner's progress to a higher level of language performance. However, the quality of learner-capable peer- interaction was crucial while scaffolding learners' development.

The implications of the study not only deliver suggestions for fairness of paired oral assessment but also provide alternatives for test-takers to achieve higher scores in a paired speaking test by choosing partners of different L2 proficiency levels. Although the outcomes did not manifest statistical significance ($p>0.05$), probably due to limited time of the interaction in the test and insufficient learning time for the test-takers, this study may provide support for relevant research in the future. The study represented a small-scale examination, and in order to have diversified groups of Chinese EFL learners and to acquire more general outcomes, it is necessary to select subjects from wider areas in China. Moreover, test-takers' characteristics were not included in the present study, and more factors can be examined in the future to analyze the underlying causes of test-takers' different performances.

Acknowledgments

The authors thank the support of the Teaching Research Project provided by Anhui Provincial Department of Education (grant No. 2021jyxm1713) and the New Liberal Arts Project provided by the University of Science & Technology of China (grant No. XWK2019015). We are also grateful to all those who participated in and supported this research, especially the participating students.

References

Bennett, R. (2011). Is linguistic ability variation in paired oral language testing problematic? *ELT Journal, 66*(3), 337-346. https://doi.org/10.1093/elt/ccr066

Bodrova, E., & Leong, D. (1996). *The Vygotskian Approach to Early Childhood.* Columbus, Ohio: Merrill, an Imprint of Prentice Hall.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Language Testing, 26*(3), 341-366. https://doi.org/10.1177/0265532209104666

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. J. (2002). *An Investigation of Second Language Task-based Performance Assessments*. Honolulu: University of Hawaii Press.

Chotlos, J. W. (1944). IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs, 56*(2), 75-111. https://doi.org/10.1037/h0093511

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396. https://doi.org/10.1177/0265532209104667

Donato, R. (1994). Tasks and language performance assessment. In Lantolf, J. P., & Appel, G. (Eds.), *Vygotskian Approaches to Second Language Research* (pp. 33-56). Ablex: Norwood, NJ.

Donovan, C. A., & Smolkin, L. B. (2002). Children's genre knowledge: An examination of K-5 students' performance on multiple tasks providing differing levels of scaffolding. *Reading Research Quarterly, 37*(4), 428-465. https://doi.org/10.1598/RRQ.37.4.5

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing, 32*(1), 101-120. https://doi.org/10.1177/0265532214544393

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based Language Teaching: Theory and Practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108643689

Engestrom, Y. (1996). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In Chaiklin, S. & Lave, J. (Eds.), *Understanding Practice: Perspectives On Activity And Context*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625510.004

Fleer, M. (1992). Identifying teacher-child interaction which scaffolds scientific thinking in young children. *Science Education, 76*(4), 373-397. https://doi.org/10.1002/sce.3730760404

Fleer, M. (1995). *Staff-Child Interactions: A Vygotskian Perspective.* Canberra: National Capital Printers.

Gass, S., & Varonis, E. (1986). Sex differences in NNS/NNS interactions. In Day, R. (Ed.), *Talking to learn: conversation in second language acquisition* (pp. 327-351). Rowley, MA: Newberry House.

Hammond, J. (2002). *Scaffolding Teaching and Learning in Language and Literacy Education.* Newtown, Australia: PETA.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*, 1-65.

Iwashita, N. (2001). The effect of learner proficiency on corrective feedback and modified output in nonnative-nonnative interaction in Japanese as a foreign language. *System, 29*(2), 267-287. https://doi.org/10.1016/S0346-251X(01)00015-X

Khaliliaqdam, S. (2014). ZPD, scaffolding and basic speech development in EFL context. *Procedia - Social and Behavioral Sciences, 98*, 891-897. https://doi.org/10.1016/j.sbspro.2014.03.497

Long, M., & Crooks, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*(1), 27-56. https://doi.org/10.2307/3587368

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104. https://doi.org/10.1191/0265532202lt221oa

McCafferty, S. (2013). Zone of proximal development in second language acquisition. In C. A. Chappelle (Ed.), *The Encyclopedia of Applied Linguistics* (Vol. 10, pp. 6286-6294). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1303

McCafferty, S. G. (2018). Vygotsky on Consciousness and Applications to Second Language Development. In Lantolf, J. P., Pohener, M. E., & Swain, M. (Eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development* (pp. 75-88). New York: Routledge.

Mcnamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

Moll, L. (1990). *Vygotsky and Education: Instructional Implications and Applications of Sociohistorical Psychology.* Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/CBO9781139173674

Newman, D., Griffin, P., & Cole, M. (1989). *The Construction Zone: Working for Cognitive Change in School.* Cambridge, UK: Cambridge University Press.

Ohta, A. S. (2000). Rethinking interaction in SLA: developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In Lantolf, J. P. (Ed.), *Sociocultural Theory and Second Language Learning* (pp. 21-78). Oxford: Oxford University Press.

Philp, J., Adams, R., & Iwashita, N. (2014). *Peer Interaction and Second Language Learning*. New York: Routledge. https://doi.org/10.4324/9780203551349

Richards, B. (1987). Type-token ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201-209. https://doi.org/10.1017/S0305000900012885

Rogoff, B. (1990). *Apprenticeship in Thinking.* New York: Oxford University Press.

Rogoff, B., & Wertsch, J. (1984). *Children's Learning in the "Zone of Proximal Development."* San Francisco, CA: Jossey-Bass.

Saville, N., & Hargreaves, P. (1999). Assessment speaking in the revised FCE. *ELT Journal, 53*(1), 42-51. https://doi.org/10.1093/elt/53.1.42

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogic Tasks* (pp. 167-185). London: Longman.

Skehan, P. (2018). *Second Language Task-based Performance: Theory, Research, Assessment*. New York: Routledge. https://doi.org/10.4324/9781315629766

Son, Y. (2016). Interaction in a paired oral assessment: Revisiting the effect of proficiency. *Language Testing and Assessment, 5*(2), 43-68.

Storch, N. (2001a). How collaborative is pair work? ESL tertiary students composing in pairs. *Language Teaching Research, 5*(1), 29-53. https://doi.org/10.1177/136216880100500103

Storch, N. (2001b). *An investigation into the nature of pair work in an ESL classroom and its effect on grammatical development* (Unpublished doctoral dissertation). University of Melbourne, Victoria, Australia.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119-158. https://doi.org/10.1111/1467-9922.00179

Storch, N. (2017). Sociocultural theory in the L2 classroom. In Loewen, S. & Sato, M. (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (pp. 69-84). New York: Routledge. https://doi.org/10.4324/9781315676968-5

Swain, M. (2000). The output hypothesis and beyond: mediating acquisition through collaborative dialogue. In Lantolf, J. P. (Ed.), *Sociocultural Theory and Second Language Learning* (pp. 97-114). Oxford: Oxford University Press.

Swain, M. (2001). Examining dialogue: another approach to content specifications and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302. https://doi.org/10.1177/026553220101800302

Swain, M., & Watanabe, Y. (2013). Languaging: Collaborative dialogue as a source of second language learning. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 3218-3225). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0664.pub2

Tharp, R., & Gallimore, R. (1988). *Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context.* New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139173698

Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: collaborative dialogue between adult ESL learners. *Language Teaching Research, 11*(2), 121-142. https://doi.org/10.1177/1362168806074599

Wells, G. (1999). *Dialogic Inquiry: Towards a Socio-cultural Practice and Theory of Education.* New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511605895

Wertsch, J. (1985). *Vygotsky and the Social Formation of Mind.* Cambridge, MA: Harvard University Press.