THE JOURNAL OF ASIA TEFL
Vol. 18, No. 1, Spring 2021, 23-38
http://dx.doi.org/10.18823/asiatefl.2021.18.1.2.23

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Effects of Examinees' Perceptions of Interlocutor Proficiency on Paired Oral Assessment Results

Hyun-kyu Choi
Yonsei University

Hee-Kyung Lee*
Yonsei University

This study aimed to explore the effects of examinees' perceptions of interlocutors' proficiency on their test results in paired oral tests. A group of 22 Korean English as a foreign language (EFL) high school students participated in the study. They were divided into three groups based on high, intermediate, and low English proficiency and were interviewed twice: once before the test and once after. Each examinee was paired with and tested with three interlocutors with high, intermediate, and low English proficiency levels. Raw scores from each rater were analyzed to examine potential interactions among the factors using the bias analysis from the Many-Facet Rasch Model. The results showed that the examinees' perceptions about the effect of interlocutor proficiency on their test results were not consistent nor reliable enough to influence the test results. This study provides evidence that the paired format can be used as a valid and reliable tool to assess EFL learners' speaking ability regardless of interlocutor proficiency levels, which can be assumed to influence examinees' perceptions of the oral test results.

Keywords: **paired oral assessment, interlocutor proficiency, examinee perception, bias analysis**

Introduction

An interview format assessing examinees individually has been the standard since oral assessments were first introduced in the 1950s (Csépes, 2009). The interview format, however, has been criticized to a great extent because individual testing is time-intensive. Moreover, this test format does not reflect real--life communication, giving the interviewer greater influence over the examinee (Bachman, 1988; Lazaraton, 1992; Van Lier, 1989). According to Van Lier, real-life communication involves face-to-face interaction, spontaneity, and a potentially equal distribution of rights and duties in conversation, whereas in the interview format the interviewer controls the interaction to a large extent, focusing on successful elicitation of language without taking successful conversation into account. Considering the disadvantages of individual interviews as a testing method, a paired test format using authentic tasks has been proposed as an alternative (Luoma, 2004). However, the paired format may also be problematic from a measurement perspective due to the interlocutor effect (Csépes, 2009; Davis, 2009). The paired format consists of examinees interacting with each other to perform a task while an examiner observes

* Hyun-kyu Choi: first author; Hee-Kyung Lee: corresponding author

their performances and rates their language proficiency. The proficiency of both examinees is assessed at the same time during the interaction, and the examiner is usually an English teacher. Examinees' outcomes and assessment results can be affected positively or negatively by the peer examinee (Csépes, 2009). This may potentially undermine test validity, which is the ability of a test to measure what it is intended to measure (Alderson, Clapham, & Wall, 1995).

Several researchers have explored various interlocutor factors in paired testing, including familiarity (Foot, 1999; Norton, 2005), personality (Berry, 2007), proficiency (Davis, 2009; Iwashita, 1996), native language (L1) (Jenkins, 1997), background culture (Young & Halleck, 1998), and gender (Norton, 2005; O'Loughlin, 2002). Among these factors, the peer interlocutor's proficiency level has been argued as the most influential, particularly in the EFL secondary school context. Some factors, such as familiarity, age, L1, and background culture, do not need to be taken into consideration in the EFL classroom assessment context, as classmates are generally familiar with one another, are close in age, have the same L1, and have similar cultural backgrounds. Bennett (2011) argued that examinees' greatest concern was the proficiency of their paired candidate, thinking that this could affect their own speaking performance.

In actual school contexts, it is often observed that students express fear or dissatisfaction of being paired with a competent (or incompetent) partner who can affect their performance and test results negatively. Only a few studies, however, have explored the interlocutor proficiency effect in paired oral tests, and there is still a need to find a consensus among researchers as to how speaking performances are affected by an interlocutor's proficiency level (Csépes, 2009; Davis, 2009; Foot, 1999; Iwashita, 1999; Norton, 2005). This study aims to explore how interlocutor proficiency influences examinees' perceptions of their oral test results and how examinees' actual test scores are related to their perceptions of the oral test results.

Literature Review

Validity of the Paired Oral Test Format

For the last two decades, interactional competence has been at the center of discussion in speaking assessment. Since Kramsch (1986) defined interactional competence as a "dynamic process of communication built through the collaborative effort of the interactional partners" (p. 386), the collaborative effort to co-construct conversations made by interlocutors has been considered as a key element of interactional competence (He & Young, 1998; Jacoby & Ochs, 1995).

Interactional competence is composed of two sets of features: an overall interaction quality as a macro-level reflected by extent of collaboration and task completion, and use of a wide range of interaction features as a micro-level. Wang (2015) further investigated these levels of interactional competence and found that interactional competence can also be operationalized with respect to interaction features and patterns. In terms of verbal and nonverbal interaction features, Vo (2019) found that both nonverbal and verbal interaction features contribute to the construct of interactional competence in her research comparing those features between interviews and paired-format speaking test. The authors all emphasized that in speaking assessment the paired format is more valid for testing examinees' interactional competence.

In addition, other researchers supported the use of a paired oral test format in other aspects (Együd & Glover, 2001; Johnson, 2001; Saville & Hargreaves, 1999; Swain, 2001; Taylor, 2000; Van Lier, 1989; Young & Milanovic, 1992). For example, this format may minimize some of the limitations of an interview-based format, such as the power differential between interviewers (usually teachers) and interviewees (usually students) and the question-and-answer style of discourse that may not accurately reflect real-life conversation (Johnson, 2001; Van Lier, 1989; Young & Milanovic, 1992). Oral communication between peers in the paired test format is similar to communication in daily life, contributing to the assessment of real-life 'target tasks' (Long & Crookes, 1992; Norris, 1998). Paired

testing may also include more types of communication than the traditional interview and, thus, increases the amount of evidence that can be gathered regarding the examinees' speaking skills (Swain, 2001); in this way, a greater variety of language functions can be utilized to produce more comprehensive samples of spoken language (Skehan, 2001). In addition, testing students in pairs may motivate students to collaborate more, which can stimulate the development of their proficiency (Saville & Hargreaves, 1999; Taylor, 2000). Most importantly, the paired test format is economical. It reduces the amount of examination time needed for testing because two examinees are assessed simultaneously (Swain, 2001). In practice, this is beneficial in EFL contexts, where class sizes are relatively large.

Various Interlocutor Effects in Paired Speaking Assessments

Although the paired format of speaking assessment has several advantages in terms of authentically measuring speaking ability, there are still questions about the effect of an interlocutor on an examinee's performance. Katona (1998) investigated the effect of examinees' familiarity with interviewers; the interviewers were paired with 12 English as a Second Language (ESL) examinees in Hungary, and two situations were compared: a practice test with a familiar teacher and a live exam with two unfamiliar examiners. Katona argued that familiarity affected the frequency and type of negotiation in the exchanges between the interviewers and candidates. O'Sullivan (2002, 2008) also examined the effect of familiarity on scores earned by pairs of Japanese English language learners. According to O'Sullivan, in the Japanese context, the degree of a test taker's familiarity with his or her interlocutor affected their performance in terms of both accuracy and complexity. Norton (2005) reported similar results in a discourse sample of Cambridge speaking tests; examinees paired with friends seemed more relaxed and performed better than examinees paired with strangers.

Furthermore, an interlocutor's personality has been suggested as a factor that may influence examinee performance. Berry (2007) divided a group of 32 male and 22 female students into two groups, introverts and extroverts, and tested them in individual interviews and paired collaborative tasks. Berry found that on individual tasks, the introverts performed better, while on the paired tasks, the extroverts performed better, regardless of the pairing type (i.e., homogeneous or heterogeneous). He also found that the introverts performed significantly better when paired with an extrovert. Bonk and Van Moere (2004) investigated the personality effect, particularly the relationship between shyness and outgoingness in speaking performance, and found that shyness affected the test scores negatively, whereas outgoingness affected them positively.

Other interlocutor effects on scores, such as L1, background culture, and gender, were also examined. Jenkins (1997) found that examinees performed better with partners sharing the same L1 due to the absence of pronunciation difficulties, which otherwise significantly hindered communication between examinees and partners. Regarding background culture, Young and Halleck (1998) found that Japanese examinees spoke more slowly and shifted topics less frequently than Mexican examinees, affecting their oral performance tests. O'Loughlin (2002) studied the effect of gender on the interview format and found that an interlocutor's gender did not affect examinees' performance significantly. However, the gender effect is closely related to sociocultural norms, which could influence a test taker's oral performance. Gender is still regarded as an unpredictable factor in testing processes and outcomes (Brown & McNamara, 2004).

Effects of Interlocutors' Proficiency

The effects of interlocutors' proficiency in paired speaking tests have been investigated in several studies since the paired format was introduced in 1939 in the First Certificate in English (FCE). Several researchers have raised concerns about varying ability levels in paired examinees (Csépes, 2009; Davis, 2009; Foot, 1999; Iwashita, 1999; Norton, 2005; Teng, 2014) and reported mixed study results. Foot (1999) found that interlocutors' proficiency might have problematic effects on comprehension in

interlocutor-examinee communication. Iwashita (1999) argued in a study of 24 students learning Japanese as a foreign language that in mixed level (high-low) dyads, more negotiation occurred when a learner was paired with a high proficiency learner compared to pairings of same-level learners. It was also found that learners' speech output increased when paired with a high proficiency partner, but scores were not affected in such pairings. Norton (2005) found similar results regarding the interlocutor proficiency effect. According to Norton, pairing potentially affects linguistic performance if one candidate has a higher linguistic ability than the other. Specifically, low proficiency examinees tended to earn higher scores on paired oral tests when paired with high proficiency examinees. That means examinees with low language proficiency were able to benefit from partners with higher language proficiency during paired tests. Teng (2014) also found that low proficiency students performed significantly better in fluency, earning higher scores and producing more language, when paired with high proficiency students, while high proficiency students performed worse when paired with low proficiency students.

On the contrary, Csépes (2009) explored the interlocutor proficiency effect in a study of 120 paired examinees in a Hungarian school, revealing no statistically significant differences between the scores earned by examinees matched with high or low proficiency partners. That is, there was no peer partner effect related to differing proficiency levels. In line with this, Davis (2009) examined the influence of interlocutor proficiency on the speaking performance of a group of 20 students at a Chinese university and found that interlocutor proficiency level had no observable effect on scores; however, lower-level examinees produced more words when paired with a higher-level partner. Moreover, in an Italian context, Bennett (2011) examined examinees' perceptions regarding whether test results would be positively or negatively affected by candidate pairing. In five of the 12 tests, the examinees were paired with a same-level candidate, and in seven of the tests they were paired with a candidate of a different level. Bennett examined the participants' perceptions using pre- and post-questionnaires and argued that while most examinees believed that their test results would be affected by their candidate pairing, no evidence was found to support their beliefs, regardless of the proficiency levels of their partners.

As seen in the above studies, the effects of interlocutor proficiency are controversial and need additional research. There has not yet been any study on how examinees themselves believe that their test scores would be affected by an interlocutor's proficiency or how their perceptions of the test results affect the actual test scores. The answers to these questions may provide an explanation for the contradictory results of previous studies on the effect of interlocutor proficiency. The present study used the Many-Facet Rasch Model (MFRM) to examine the interactions among the proficiency levels of interlocutors, examinees' perceptions of their test results, and actual examinee test scores on paired oral tests. The research questions of this study are as follow:

1. How are the examinees' perceptions of the oral test results influenced by the speaking proficiency level of an interlocutor in a paired oral test? How do their perceptions differ before and after the paired oral test?
2. How are the examinees' oral test results related to the examinees' perceptions of the oral test results?

Methods

Participants

The participants in the study included examinees, interlocutors, and raters. The 22 examinees were students taking oral tests with partners of three different proficiency levels (high, intermediate, and low). They were all second-year male students in the same class at a high school in Seoul, Korea. They took the TOEIC Speaking Test (TST) to assess their current level of speaking proficiency and were then grouped according to proficiency level (high, intermediate, or low). On the TST, five students (22.73%) scored at

levels 7-8, and they were classified as the high proficiency group; 14 students (63.63%) scored at levels 4-6 and were categorized as the intermediate proficiency group, and three students (13.64%) scored at levels 1-3 and classified as the low proficiency group. The interlocutors were examinees' test partners, and three interlocutors for the 22 students (one interlocutor in each proficiency group) were selected based on the results of TST. The high proficiency interlocutor received a maximum score of 200 (level 8), the intermediate proficiency interlocutor scored 80 (level 4), and the low proficiency interlocutor scored 40 (level 2). The raters were all native English teachers serving as examiners. They had taught English to middle or high school students and evaluated their speaking abilities in school settings for at least five years. As seen in table 1, the reliability for each scoring dimension was confirmed via the 1st (before rater training) and 2nd (after rater training) correlation analysis using SPSS version 18.0.

TABLE 1
Correlation Analysis Among Three Raters After Rater Training

Category	Raters	Pearson Correlation	Sig. (2-tailed)
Grammar & Vocabulary	M & T	.772	.005
	T & N	.723	.012
	N & M	.881	.000
Pronunciation	M & T	.883	.001
	T & N	.528	.095
	N & M	.796	.003
Fluency	M & T	.912	.000
	T & N	.946	.000
	N & M	.952	.000
Discourse Management	M & T	.758	.007
	T & N	.788	.004
	N & M	.935	.000

Note. Correlation is significant at the .05 level (2-tailed).

Interview Questions

To observe the examinees' perceptions before and after the paired oral tests, semi-structured interviews were conducted twice for each examinee. A semi-structured interview allows an interviewer to ask additional questions based on the interviewee's responses to obtain further information. The predetermined question was the only one that was the same in the pre- and post-interviews:

> When taking an oral test three times with high, intermediate, and low proficiency partners, how do you think your partner's proficiency affects the test results?

When conducting interviews in Korean, the researcher asked each examinee the prepared question regarding their perceptions of the oral test results. Each examinee was interviewed twice: once before the test and once after. After concluding the interviews, the researcher transcribed the interview data. Two coders, a Korean English teacher from a school and the researcher, independently read and coded the data. The examinees' responses were coded as numbers ranging from one to five. When discrepancies were identified between coders, the coders asked the examinee the question again to ascertain exactly how they perceived the influence of the interlocutor's proficiency. A score of 1 indicates that an examinee perceives that the proficiency of an interlocutor would affect the test result 'very negatively (VN);' 2 indicates 'negatively (N);' 3 indicates 'no [zero] effect (Z);' 4 indicates 'positively (P);' and 5 indicates 'very positively (VP).'

Speaking Sample Collection

Each interlocutor engaged in oral tests with 21 examinees, and each examinee took the paired oral test three times. Each examinee tested in English with a high, intermediate, and low proficiency interlocutor (see Appendix A), and all oral performance samples were scored by three raters. When the paired oral test started, one examinee and one of three interlocutors entered the classroom and were seated with a native English-speaking teacher. The teacher randomly chose one of 21 'would you rather' question cards and gave the card to the examinee and the interlocutor, providing one minute of preparation time. When taking the oral tests, one interlocutor used 21 different 'would you rather' question cards with 21 examinees to minimize any practice effect (which would have contributed to a significant improvement in performance). After one minute, the examiner (the teacher) provided three minutes for discussing and answering the question. To avoid interrupting the conversation, the examiner did not add any further questions while the examinee and interlocutor discussed the question. The performances were video- and audio-recorded. After completing the tests, two other native English-speaking raters scored the speaking performances of all examinees using the video-taped materials and an analytical rubric (see Appendix B).

Data Analysis

Many-Facet Rasch Model (MFRM) was adopted by using FACETS version 3.71 (Linacre, 2014). In the various analyses provided in the FACETS program, multiple bias analyses were performed to examine potential interactions among the various facets: (a) interlocutors' proficiency and examinees' perceptions in the pre-interviews; (b) interlocutors' proficiency and examinees' perceptions in the post-interviews; (c) examinees' perceptions in the pre-interviews and criteria, such as grammar and vocabulary, pronunciation, fluency, and discourse management (content); (d) examinees' perceptions in the post-interviews and criteria; and (e) interlocutors' proficiency and criteria. The bias analysis can identify statistically significant interactions among the elements of different facets, evaluating the validity of ratings by examining systematic interactions between elements of a certain facet and those of other facets (Linacre, 2014). In the bias analysis, the significance (*p*-value) is important in that it indicates whether the fixed hypothesis is true and that the set of interactions is considered to share the same measure (about 0.0) after allowing for measurement error (Linacre, 2014). If a bias analysis as performed by a fixed chi-square test is significant, it suggests that the overall magnitude of the bias is too large to result from chance (Aryadoust, 2016). The *t*-value is also important in that it provides information about whether each bias magnitude is statistically significant, ensuring that there is no bias except for measurement error (Linacre, 2014). Any *t*-values outside of the +/-2 range suggest significant bias. The fit statistics determine whether the magnitude of bias might have been overlooked. When analyzing the data, one type of fit statistic is often used, and infit statistics are preferred over outfit statistics. The acceptable range of infit and outfit mean square statistics is from .5 to 1.5. A fit statistic greater than 1.5 indicates a misfit, or unpredictability in an element of a facet, whereas a fit statistic less than .5 indicates overfit, or not enough variation in scores.

Results and Discussion

Examinees' Perceptions of Test Results Influenced by their Interlocutor's Oral Proficiency

To identify the interactions between the examinees' perceptions of their oral test results and the interlocutors' English proficiency, a bias analysis was carried out. Table 2 shows a bias report between interlocutors' proficiency and examinees' perceptions in the pre-interviews, which indicate examinees'

expectations regarding their test results in the paired oral tests (very positive [VP], positive [P], no (zero) effect [Z], negative [N], and very negative [VN]). The bias tables show only the significant results among the factors from the bias analyses.

TABLE 2
Interaction Analysis Between Interlocutors' Proficiency and Examinees' Perceptions in the Pre-interview

Interlocutor's proficiency level	Examinee's perception	Obs-Exp Average	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
High	VP	-.53	-1.22	.25	-4.68	35	.00	1.3
High	N	.15	.38	.16	2.42	107	.02	1.0
Low	VN	.33	.83	.27	3.06	35	.00	1.0
Fixed (all = 0) chi-square: 42.0			*d.f.*: 13		significance (probability): .00			

As shown in Table 2, the fixed chi-square test for the bias analysis was significant ($\chi^2 = 42.0$, $d.f. = 13$, $p < .05$), suggesting that the magnitude of bias between the interlocutors' proficiency and the examinees' perceptions of their oral test results in the pre-interviews was generally too large to be due to chance. This suggests that there were bias interactions between the interlocutors' proficiency and the examinees' perceptions of their oral test results in the pre-interviews; the examinees' expectations of their oral test results before taking the tests were systematically affected by the interlocutors' proficiency. This concurs with the findings regarding the effects of interlocutors' proficiency in paired oral tests expressed by Bennett (2011). According to Bennett, among 43 respondents to a pre-test questionnaire, only three believed that their performance would not be affected by the interlocutor, and the greatest concern of the remaining 40 was the proficiency of the paired candidate.

To examine the bias interactions between the interlocutors with different proficiency levels and the examinees' perceptions of their oral test results, a *t*-statistic was performed. As shown in table 2, three bias interactions were found between the interlocutors' proficiency and the examinees' perceptions in the pre-interviews, as determined by the *t*-statistic and the infit mean-square probability. The bias size between the interlocutors with high proficiency and the examinees' perception of their oral test results in the pre-interviews, very positive, was found to be -1.22 logits, negatively biased, which is statistically significant ($t = -4.86$, $p < .05$). Moreover, the infit mean-square was 1.3, which is lower than 1.5 and higher than .5, indicating that the interactions were consistent. There were examinees who expected their oral test results to be affected very positively when they took paired oral tests with high proficiency interlocutors. On average, they were awarded .53 lower oral test scores than they expected. This is shown as 'Obs-Exp Average' in table 2.

On the other hand, the bias size between the highly proficient interlocutors and the examinees' perception of their oral test results, negative, was .38 logits, positively biased, which is statistically significant ($t = 2.42$, $p < .05$). Moreover, the infit mean-square was 1.0. There were examinees who expected that their oral test results would be affected negatively when they took paired oral tests with interlocutors with high proficiency. On average, they received .15 higher oral test scores than they expected. Similarly, the bias size between the interlocutors with low proficiency and the examinees' perception of their oral test results, very negative, was .83 logits, positively biased, which is statistically significant ($t = 3.06$, $p < .05$), and the infit mean-square was 1.0. There were examinees who expected their oral test results to be affected very negatively when they took paired oral tests with low proficiency interlocutors. On average, they received .33 higher oral test scores than they expected.

Through the previous analysis of bias interactions, however, it is impossible to recognize which examinee English proficiency levels show systematic bias in expecting how their oral test results would be affected. Accordingly, a bias analysis was carried out again for the interactions of the interlocutors' and examinees' proficiency, and the examinees' perceptions in the pre-interviews. As shown in table 3, the fixed chi-square test was significant ($\chi^2 = 55.8$, $d.f. = 24$, $p < .05$), suggesting that the magnitude of the bias was generally too large to be due to chance. This suggests that there were overall bias interactions among the interlocutors' and examinees' proficiency, and the examinees' perceptions of their oral test

results in pre-interviews. Specifically, three bias interactions were found among the interlocutors' and examinees' proficiency and the examinees' perceptions in the pre-interviews. One thing to note is that the examinees who perceived that their oral test results would be affected very positively or (very) negatively by the interlocutors had intermediate proficiency levels. This suggests that in the pre-interviews, the perceptions of high and low proficiency examinees were not affected by the interlocutors, regardless of whether the interlocutors were highly proficient or not; on the other hand, the perceptions of the intermediate proficiency examinees were affected by the interlocutors, especially with high and low proficiency interlocutors.

TABLE 3
Interaction Analysis Among Interlocutors' Proficiency, Examinees' Proficiency, and Examinees' Perceptions in the Pre-interview

Interlocutor's proficiency level	Examinee's proficiency level	Examinee's perception	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
High	Intermediate	VP	-1.22	.25	-4.86	35	.00	1.3
High	Intermediate	N	.45	2.64	2.64	83	.01	1.0
Low	Intermediate	VN	.83	3.06	3.06	35	.00	1.0
Fixed (all = 0) chi-square: 55.8			*d.f.*: 24		significance (probability): .00			

Table 4 shows bias report between the interlocutors and the examinees' perceptions in the post-interviews, and one bias interaction was found. The bias size between the interlocutors with high proficiency and the examinees' perception of their oral test results in the post-interviews, very negative, was -1.28 logits, negatively biased, which is statistically significant (t = -2.93, $p < .05$). However, there was no overall systematic bias interaction between the interlocutors' proficiency and the examinees' perceptions of their oral test results in the post-interviews. The fixed chi-square test was not significant (χ^2 = 17.7, $d.f.$ = 12, $p > .05$), suggesting that the magnitude of bias was generally small enough to be due to chance. There was no overall statistically significant bias interaction between the interlocutors' proficiency and the examinees' perceptions of their oral test results in the post-interviews. This suggests that after a test, the examinees believed that their performance had not been positively or negatively affected by the interlocutor's proficiency. This result differs from previous research on the effects of interlocutors' proficiency in paired oral tests (Bennett, 2011). According to Bennett, after the tests, none of the examinees believed their performance with partners of different levels had been adversely affected, whereas 50 percent believed their performance had been enhanced, irrespective of whether they were tested with a high- or low-level partner.

TABLE 4
Interaction Analysis Between Interlocutors' Proficiency and Examinees' Perception in the Post-interview

Interlocutor's proficiency level	Examinee's perception	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
High	VN	-1.28	.44	-2.93	11	.01	.4
Fixed (all = 0) chi-square: 17.7		*d.f.*: 12		significance (probability): .12			

The findings above suggest that the examinees' perceptions of the oral test results depended on whether the examinee was questioned before or after the test. The examinees' perceptions of their test results changed while taking the paired oral tests. In other words, prior to the oral tests, the examinees expected that their test results would be affected by the interlocutors' proficiency, whereas after the tests, the examinees thought that their results were not influenced by proficiency. This demonstrates that the examinees' perceptions of their test results were not consistent or reliable. This result could be regarded as supportive of paired oral testing in that the method could have validity regardless of examinees' perceptions.

Examinees' Perceptions of the Oral Test Results Compared to Their Actual Test Results

To identify the interactions between the examinees' perceptions of their oral test results before the oral tests and their actual oral test results, a bias analysis was performed. Table 5 shows a bias report between the examinees' perceptions of their oral test results beforehand along with their actual oral test results, and the fixed chi-square test was significant ($\chi^2 = 34.8$, *d.f.* = 20, $p < .05$), suggesting that the magnitude of bias was generally too large to be due to chance. This suggests that there were overall bias interactions between the examinees' perceptions in the pre-interviews and the criteria. That is, the examinees' expectations of their oral test results before taking the tests had systematic interactions with their actual test results.

To examine the bias interactions between the examinees' perceptions in the pre-interviews and the criteria such as fluency, discourse management, pronunciation, and grammar and vocabulary, it was necessary to examine whether the *t*-statistic is significant. In Table 5, three bias interactions are shown. The bias size between the examinees' perceptions of their oral test results in the pre-interviews, very positive, and the result in the 'discourse management' domain, was -1.05 logits, negatively biased, which is statistically significant ($t = -2.41$, $p < .05$); the infit mean-square was .9. Similarly, the bias size between the examinees' perceptions of their oral test results in the pre-interviews, very positive, and the result of the 'pronunciation' domain, was -.98 logits, negatively biased, which is statistically significant ($t = -2.25$, $p < .05$); the infit mean-square was 1.0. On the other hand, the bias size between the examinees' perceptions of their oral test results, very negative, and the result in the 'discourse management' domain, was .98 logits, positively biased, which is statistically significant ($t = 2.40$, $p < .05$). Moreover, the infit mean-square was 1.0.

TABLE 5
Interaction Analysis Between Examinees' Perceptions in the Pre-interview and Criteria

Examinee's perception	Criteria	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
VP	D	-1.05	.44	-2.41	11	.03	.9
VP	P	-.98	.44	-2.25	11	.04	1.0
VN	D	.98	.41	2.40	17	.03	1.0
Fixed (all = 0) chi-square: 34.8		*d.f.*: 20		significance (probability): .02			

Note. Criteria: D (discourse management), P (pronunciation)

Through the previous analysis of bias interactions, it was impossible to ascertain which proficiency levels of examinees showed systematic bias in expecting how their oral test results would be affected, especially in the specific criteria domains. Accordingly, a bias analysis was carried out for the interactions of the examinees' proficiency, the pre-interviews, and the criteria. Table 6 shows that the fixed chi-square test was significant ($\chi^2 = 92.7$, *d.f.* = 48, $p < .05$), suggesting that the magnitude of bias was generally too large to be due to chance, which suggests that there were overall bias interactions. Specifically, two bias interactions were found among the examinees' proficiency, their perceptions in the pre-interviews, and the criteria. One thing to note is that the proficiency level of examinees who perceived that their oral test results would be affected very positively was intermediate. This suggests that the perceptions of the high and low proficiency examinees in the pre-interviews were not consistent with their test scores, while the perceptions of the intermediate proficiency examinees corresponded with their test scores.

TABLE 6
Interaction Analysis Among Examinees' Proficiency, Perceptions in the Pre-interview, and Criteria

Examinee's proficiency level	Examinee's perception	Criteria	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
Intermediate	VP	D	-1.05	.44	-2.41	11	.03	.9
Intermediate	VP	P	-.98	.44	-2.25	11	.04	1.0
Fixed (all = 0) chi-square: 92.7			*d.f.*: 48		significance (probability): .00			

Table 7 shows a bias report between the examinees' perceptions in the post-interviews and the criteria. Generally, there was no systematic bias interaction. The fixed chi-square test was not significant (χ^2 = 19.5, *d.f.* = 20, $p > .05$), suggesting that the magnitude of the bias is generally small enough to be due to chance. This suggests that there was no overall bias interaction between the examinees' perceptions in the post-interviews and the criteria.

TABLE 7
Interaction Analysis Between Examinees' Perceptions in the Post-interview and Criteria

Examinee's perception	Criteria	Bias Size	Model S.E.	*t*	*d.f.*	*p*	Infit MnSq
Fixed (all = 0) chi-square: 19.5		*d.f.*: 20		significance (probability): .49			

In asking whether the examinees' oral test results were influenced by the oral proficiency level of the interlocutor, the answer is no. No bias interaction was found between the interlocutors' proficiency and the criteria, and there was no overall systematic bias interaction. The fixed chi-square test was not significant (χ^2 = 4.6, *d.f.* = 12, $p > .05$), suggesting that the magnitude of bias was generally small enough to be due to chance. The proficiency levels of the interlocutors appeared to have no statistically significant interaction with the test results of the examinees, and it can be said that the test scores of the examinees were not positively or negatively affected by the proficiency levels of the interlocutors. This concurs with the findings in studies conducted by Bennett (2011), Davis (2009), and Iwashita (1999). According to Bennett, contrary to examinees' beliefs, differences in interlocutors' linguistic abilities had little effect on scores. Davis also argued that no statistically significant differences were observed in raw scores when examinees were paired with either high or low proficiency partners. Iwashita revealed that the quantity of speech was affected when paired with a high proficiency partner, but the examinee's score was not impacted.

The findings above suggest that the examinees' oral test results were not related to their perceptions of the oral test results. Overall, prior to the tests, the examinees' perceptions of their results had a systematic interaction with their actual test results, whereas after the tests the examinees' perceptions had no systematic interactions with actual test results; also, the proficiency levels of the interlocutors did not affect the test results of the examinees, regardless of the examinees' perceptions. It seemed that the examinees' perceptions of their test results prior to the tests could be attributed to anxiety and nervousness, which could have caused them to believe that their performance would be severely affected by external environmental factors. However, soon after the test, the anxiety and nervousness would have turned into relief, allowing them to focus solely on their own performance. This demonstrates that the examinees' perceptions of their test results were not trustworthy or consistent, and EFL teachers do not need to worry about the effects of interlocutors' proficiency during paired oral tests; there should be little to no problem with the use of the paired oral test format as an assessment tool in the EFL classroom context. Regardless of interlocutors' language proficiency levels, the paired format can be used as a valid and reliable tool to assess EFL learners' speaking abilities.

Conclusion

This study explored how interlocutors' proficiency influenced examinees' perceptions of their oral test results and whether examinees' actual test scores were consistent with the examinees' perceptions of the effects of interlocutor proficiency on their test results. Within the context of this study, the examinees' perceptions of their test results depending on the interlocutor's proficiency were not consistent or trustworthy. Before taking the tests, the proficiency levels of examinees' partners in the paired oral tests had significant influences on the examinees' expectations of the oral test results. More specifically, before the tests, the examinees did not have favorable or unfavorable expectations regarding their results when paired with an intermediate level interlocutor, but some examinees perceived that pairing with a high or a

low proficiency partner was either very favorable or very unfavorable. However, after the tests, the examinees perceived that their test results were not influenced by the interlocutors' proficiency. Interestingly, even before the tests, only the intermediate proficiency examinees' perceptions on their test scores were affected by the interlocutors' proficiency. The high and low proficiency examinees did not perceive that their oral test results would be influenced by the interlocutors' proficiency. In addition, the examinees' perceptions of their oral test results prior to the tests were congruent with their actual test results, whereas after the tests, the examinees' perceptions did not have significant interactions with their results. In the interviews prior to the tests, only the intermediate proficiency examinees perceived that their performance would be affected positively or negatively by the interlocutor's proficiency, while the other two groups did not report such perceptions.

Notably, only 22 students from one secondary school participated in the study, and compared to the number of intermediate proficiency participants, the numbers of high and low proficiency participants were small (high: 5 students; intermediate: 14 students; low: 3 students); the participants all came from a heterogeneous proficiency class. Moreover, only one type of prompt, the 'would you rather' questions, was used to assess the participants' English ability in the paired oral tests; this could have affected the authenticity and language function diversity. More valid measures of the participants' speaking abilities could have been acquired by combining several tasks (Son, 2013). Finally, the current study could provide only limited information regarding the effects of the interlocutors' proficiency in paired oral test format due to the use of quantitative analysis only. Using qualitative analysis along with the current method might provide a more comprehensive picture of the examinees' perceptions, including why the examinees perceived that their test scores would be affected and why their perceptions changed over time as the testing progressed.

Despite the limitations above, when EFL secondary students express fear or dissatisfaction and claim that their oral performance and test scores would be negatively affected by the language proficiency of their partners, classroom teachers and language assessors have a basis for reassuring students that their concerns are unwarranted and their performance is assessed solely based upon their own speaking performance. Considering the benefits of paired oral testing, including efficiency, the absence of an interlocutor proficiency effect, and the incongruence between examinees' post-test perceptions of their test results and their actual test results, the paired oral testing format can be considered a good option for English assessment in the EFL classroom context.

The Authors

Hyun-kyu Choi is a lecturer at the Graduate School of Education in Yonsei University, Seoul, Korea as well as an English teacher in Yangchung High School. His primary interests are in language measurement and assessment, learning strategies, and teaching methods.

Major of English Education
Graduate School of Education
Yonsei University
50 Yonsei-ro, Seodaemun-gu,
Seoul, 03722, Korea
Email: calebchoi11@hanmail.net

Hee-Kyung Lee is a professor at the Graduate School of Education in Yonsei University, Seoul, Korea. Her research areas are English curriculum development, second language writing, language testing, and classroom assessment.

Major of English Education
Graduate School of Education

Yonsei University
Tel: +82 2 2123 6265
Email: heelee@yonsei.ac.kr

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly, 13*(1), 1-24.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Bennett, R. (2011). Is linguistic ability variation in paired oral language testing problematic? *ELT Journal, 66*(3), 337-346.

Berry, V. (2007*). Personality differences and oral test performance* (Vol. 7). Peter Lang.

Bonk, W. J., & Van Moere, A. (2004, March). L*2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores.* Paper presented at the Language Testing Research Colloquium, Temecula, California.

Brown, A., & McNamara, T. F. (2004). The devil is in the detail. *TESOL Quarterly, 38*(3), 524-538.

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance* (Vol. 14). Peter Lang.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Együd, G., & Glover, P. (2001). Oral testing in pairs-a secondary school perspective. *ELT Journal, 55*(1), 70-76.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36-41.

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins Publishing Company.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-65.

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

Jenkins, J. (1997). Testing pronunciation in communicative exams. *Speak Out*, *20*, 7-11.

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview*. New haven, CT: Yale University Press.

Katona, L. (1998). Meaning negotiation in the Hungarian oral proficiency examination of English. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 239-270). Amsterdam: John Benjamins Publishing Company.

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

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

Linacre, J. M. (2014). A user's guide to FACETS: Rasch model computer program. Retrieved from http://www.winsteps.com/a/facets-manual.pdf/

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*(1), 27-56.

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

Norris, J. M. (1998). *Designing second language performance assessments* (No. 18). National Foreign Language Resource Center.

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

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

O'Sullivan, B. (2002). Using observation checklists to validate speaking test pair tasks. *Language Testing, 19*(1), 33-56.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language* (Vol. 7). Peter Lang.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*(1), 42-51.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *In researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167-185). London: Longman.

Son, Y. A. (2013). *The interlocutor proficiency effect on test-taker performance in a paired oral assessment* (Unpublished master's thesis). Seoul National University, Korea.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302.

Taylor, L. (2000). Investigating the paired speaking test format. *Cambridge ESOL Research Notes*, 2, 14-15.

Teng, H. C. (2014). Interlocutor proficiency in paired speaking tests. In N. Sonda & A. Krause (Eds.), *JALT 2013 Conference Proceedings*. Tokyo: JALT.

Van Lier, L. (1989). Reeling, writing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly* 23, 489-508.

Young, R., & Halleck, G. B. (1998). "Let them eat cake!" or how to avoid losing your head in cross-cultural conversations. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 355-382). Amsterdam: John Benjamins Publishing Company.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(4), 403-424.

Vo, S. (2019). *Effects of task types on interactional competence in oral communication assessment* (Unpublished doctoral dissertation). Iowa State University, USA.

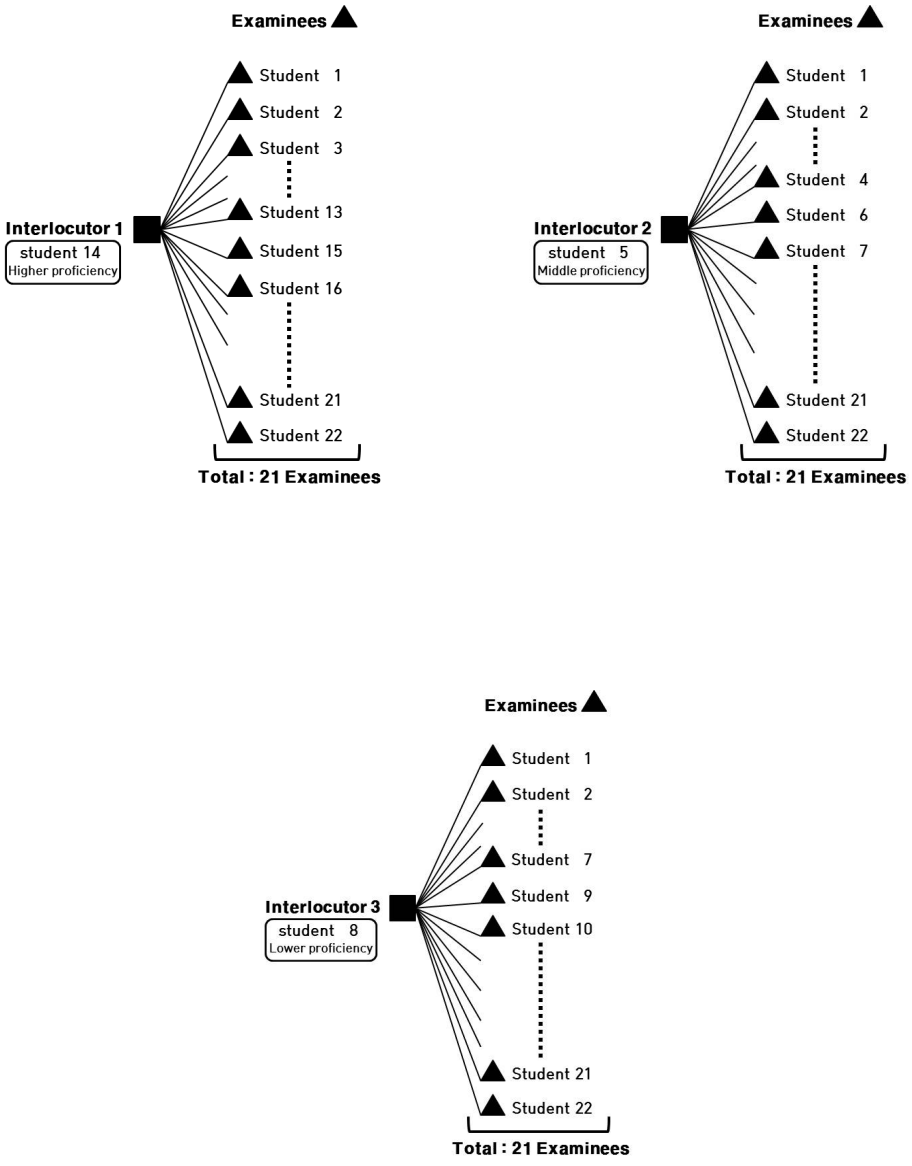
Wang, L. (2015). *Assessing interactional competence in second language paired speaking tasks* (Unpublished doctoral dissertation). Northern Arizona University, USA.

(Received January 07, 2021; Revised February 28, 2021; Accepted March 10, 2021)

The Journal of Asia TEFL
Vol. 18, No. 1, Spring 2021, 23-38

Appendix A

Interlocutor-Examinee Pairing for Speaking Tests



Appendix B

Scoring Rubric (Davis, 2009, pp. 394-395)

	Score = 1	Score = 2	Score = 3	Score = 4	Score = 5
Grammar & Vocabulary	Produces only very basic sentence forms. Overall, turns are short, structures are repetitive, and errors are frequent. Uses simple words almost exclusively; inappropriate usages common. Errors are often distracting, and may damage communication.	Primarily uses basic sentences; more complex structures are absent or contain significant errors. Vocabulary sufficient to discuss topic, but generally simple. Errors are common, and may be distracting at times.	Produces a mix of short and complex sentence forms, typically uses shorter forms. Vocabulary is adequate to discuss topics at length, although simplifications may be common. Errors in grammar and vocabulary are noticeable.	Uses a range of structures, but errors may be fairly common. Vocabulary is generally appropriate with some words being precise, although inappropriate usages are evident. May contain a number of repetitive errors.	Makes use of longer sentences and a variety of structures. Uses a range of vocabulary; words are mostly precise. Errors remain, but are not distracting.
Pronunciation	Errors in pronunciation are severe enough that some words or utterances are hard to understand or indecipherable. Intonation is generally unnatural and errors in individual sounds are very common.	Pronunciation is good enough to be understood, although there may be some difficulty in understanding at times. Errors in producing single sounds are frequent, and intonation may be noticeably unnatural. Repetitive errors may be common.	Pronunciation is good enough to be understood throughout. Able to approximate natural intonation, although unnatural intonation or errors in single sounds are present. May contain a number of repetitive errors.	Occasional errors in intonation or single sounds are noticeable, but not distracting. May contain one or two repetitive errors.	Errors in intonation or single sounds are present, but few and/or minor. Noticeable "foreign" accent is present, but not distracting.
Fluency	Unable to keep going without noticeable pauses. Long pauses are common. Speech is slow with frequent repetition.	Usually able to keep going, but relies on repetition or self-correction to do so and/or on slow speech. Hesitation is common, and may include several periods of silence.	Able to keep going and shows willingness/ability to produce long utterances. Hesitation is present, but shows some use of hesitation devices. May have a few longer periods of hesitation or periods of silence.	Generally able to keep going and readily produces long utterances. Uses appropriate hesitation devices when needed, although long periods of hesitation may occur. Uses strategies to maintain and repair interaction.	Able to keep going, but with occasional hesitation, repetition, or self-correction. Uses appropriate hesitation devices when needed, long periods of hesitation are few or absent. Actively engages with partner.
Discourse Management (Content)	Produces only very simple arguments or opinions. Breakdowns in coherence or inappropriate responses to partner's speech are noticeable.	Arguments are mostly simple, and generally does not develop detailed reasoning. May show occasional instances of unclear logic or inappropriate responses.	Reasoning is typically fairly simple, although longer sequences may occur. Generally able to respond appropriately to the task and to the partner's contribution, although somewhat superficially. May show a few inappropriacies.	Produces extended argument or opinion, although typical contributions are more simple. Utterances are generally arranged logically. Responses to partner are appropriate, although some may be fairly superficial.	Readily produces extended and/or relatively complex arguments or opinions, and makes thoughtful responses to other's arguments.

Appendix C

21 "Would You Rather" Question Samples

1. Would you rather own a dog or a cat?
   I would rather own a cat.
   I would rather own a dog.
2. Would you rather learn to dance or to box?
   I would rather learn to dance.
   I would rather learn to fight.
3. Would you rather have no arms or no legs?
   I would rather have no arms.
   I would rather have no legs.
4. Would you rather be president or the pope?
   I would rather be president.
   I would rather be the pope.
5. Would you rather be rich or immortal?
   I would rather be rich.
   I would rather be immortal.
6. Would you rather be able to fly or turn invisible?
   I would rather be able to fly.
   I would rather be able to turn invisible.
7. Would you rather be deaf or blind?
   I would rather be deaf.
   I would rather be blind.
8. Would you rather have no toothpaste forever or no shampoo forever?
   I would rather have no toothpaste.
   I would rather have no shampoo.
9. Would you rather be a child for a day or the opposite gender for a day?
   I would rather be a child for a day.
   I would rather be a woman/man for a day.
10. Would you rather master a foreign language or a musical instrument?
    I would rather master a foreign language.
    I would rather master a musical instrument.

