The Comparability of Direct and Semi-direct Oral Proficiency Interviews in a Foreign Language Context: A Case Study with Advanced Korean Learners of English

Ikkyu Choi*
(University of California, Los Angeles)

Choi, Ikkyu. (2014). The Comparability of Direct and Semi-direct Oral Proficiency Interviews in a Foreign Language Context: A Case Study with Advanced Korean Learners of English. ***Language Research,*** **50.2, 545-568.**

The comparability between oral proficiency interviews (OPI) and semi-direct oral proficiency interviews (SOPI) has been an important issue since the advent of the SOPI as an alternative to the OPI. This study examines whether the two test modes can be considered to elicit comparable performance in a foreign language test context. 15 advanced Korean learners of English were given two oral proficiency tests consisting of a matched set of tasks: one under a simulated OPI and the other under a simulated SOPI. The resulting test scores and performance samples were compared. The results showed that the scores and the performance samples from the two test modes were highly comparable. This comparability is an encouraging result since the SOPI is more practical to administer than the OPI, especially in a foreign language context.

Keywords: oral proficiency interview, semi-direct oral proficiency interview, English oral proficiency, interviewer effect

Ⅰ. Introduction

Improving oral proficiency is an important goal of many curricula in foreign and second language teaching. Similarly, oral proficiency has been one of the most popular constructs of foreign and second language assessments. Many attempts have been made to properly measure one's oral proficiency. Early oral proficiency measures often employed mechanical repetitions and substitution drills as their main elicitation tech-

* The author is now at Educational Testing Service.

niques (Shohamy, Reves, & Bejerano 1986). However, due to the growing demands for direct assessment of one's communicative ability, these earlier elicitation techniques were deemed unsatisfactory. Consequently, many professionals in the field of language teaching and testing began to develop oral proficiency assessments that can measure test takers' communicative ability in a more direct and effective manner. Well-known products of such efforts include oral proficiency interviews (OPI) and semi-direct oral proficiency interviews (SOPI).

The OPI was developed by the United States Foreign Service Institute (FSI) and has gained wide popularity since its introduction in the 1950s. Employing a series of interviews between interviewers/raters and interviewees/test takers, the OPI was designed to simulate characteristic features of real-life language use situations in language testing contexts. The SOPI, on the other hand, elicits and evaluates performance samples by means of pre-recorded audio/video stimuli, without the presence of a live interviewer. Despite its less direct nature, the SOPI has been widely regarded as a more practical and cost-efficient alternative of the OPI. The practicality of the SOPI is especially pertinent in foreign language testing situations in which it is difficult to establish a reliable pool of qualified interviewers. Furthermore, the SOPI makes it possible to ensure the same elicitation procedure across all test takers, and therefore, is free from concerns about the heterogeneous interviewer effect that is inherent in the OPI (Brown 2003). Therefore, some researchers (e.g. Lazaraton 1996; Stansfield 1990) regard the SOPI as a more reliable measure of oral language proficiency than the OPI.

The compatibility between the OPI and the SOPI has been an important issue since the advent of the latter as an alternative to the former (e.g., Stansfield & Kenyon 1992; Wigglesworth & O'Loughlin 1993). The SOPI is attractive thanks to its practical and standardized nature, but lacks the appeal of the OPI, which directly simulates a face-to-face communication. Therefore, if the SOPI is to be considered as a more practical and standardized alternative to the OPI, it is of central importance to examine whether the two test modes elicit comparable information from test takers.

The comparability of the two test modes is not free from the context of an assessment. While there have been several empirical investigations of their comparability in second language environments (e.g., O'Loughlin 2001; Shohamy 1994; Stansfield & Kenyon 1992), little efforts to directly compare language samples elicited from the two test modes have been made in foreign language contexts. To bridge this research gap, this study aims to gather empirical evidence of the comparability between the two test modes in a foreign language testing context. In particular, this study compares the scores and the performance samples obtained under the two test modes from advanced Korean learners of English. The results of the comparison can provide empirical grounds for making decisions in designing and administering an oral proficiency test in a foreign language context.

Ⅱ. Review of Literature

Early investigations of the comparability between the OPI and the SOPI primarily, and in some cases exclusively, relied on correlation-based approaches. Clark (1979) maintained that a correlation of .9 or higher would be required to justify the use of SOPI results as probable alternatives of the results from more direct measures. Clark and Swinton (1980) argued that, based on the correlation of .8 between the scores obtained from international teaching assistants, the Test of Spoken English (TSE), a semi-direct type oral proficiency test, could be reasonably accepted as an alternative to the FSI interview. A series of studies (e.g., Shohamy et al. 1989) were conducted during the development processes of the SOPI in various languages, and they also found high correlations between the scores from the OPI and the SOPI. Similarly, Stansfield and Kenyon (1992) claimed that the OPI and the SOPI were highly comparable means of assessing one's speaking ability based on score comparisons.

As the correlation-based evidence for the comparability of the OPI and SOPI scores continued to accumulate, researchers have begun to investigate the comparability from multiple perspectives. Shohamy (1994) administered the OPI and the SOPI for Hebrew and compared the resulting performance samples as well as the scores from the two tests. Adopting a series of qualitative analyses, she found that, despite high correlations between the scores, the performance samples elicited under the two test modes differed in terms of characteristic discourse features. The performance samples from the SOPI were, in general, more 'literate' than the performance samples from the OPI, which resembled oral conversations more closely. Shohamy further claimed that the test mode (i.e., the OPI or the SOPI) could have stronger influence on elicited oral language than test tasks.

O'Loughlin (2001) extended Shohamy's (1994) approach in his comparison of the OPI and SOPI versions of the *access:* test. His results, however, only partially concurred with Shohamy's results. In particular, the two versions yield somewhat different test taker ability estimates, while most discourse features in the performance samples elicited from the two versions were highly comparable within the same task type. O'Loughlin attributed the discrepancy of his results from Shohamy's results to differences in their study design. While Shohamy compared elicited performance samples directly from the Hebrew OPI and SOPI tests without controlling for potential task effects, O'Loughlin controlled for task effects in his OPI and SOPI versions by providing matched tasks. Therefore, in O'Loughlin's design, task effects were not confounded with potential mode effects. In addition, the OPI raters/interviewers in O'Loughlin's study were instructed to keep their interaction with test takers to a minimum, whereas Shohamy did not attempt to control the interviewer-test taker interaction.

Both Shohamy (1994) and O'Loughlin (2001) examined the comparability of the OPI and the SOPI in second language contexts. The two oral proficiency test modes have also attracted interest from researchers working in foreign language contexts. H Jung (2000) and I-C Choi (2000) investigated the validity of tests that adopted the SOPI. D-I Shin and J-K Kim (2005) adopted discourse analytic approaches to study performance samples elicited under an OPI setting. More recently, the comparability of the OPI and the SOPI was examined from the perspectives of stakeholder. Qian (2009) studied the popularity of the OPI and the SOPI among university students in Hong Kong. M-J Joo (2007) also investigated the attitude toward the OPI and the SOPI of English learners and teachers at a Korean university. Moreover, Jeong et al. (2011) compared test takers' neural processes during OPI and SOPI administrations. However, a direct comparison of test taker performance elicited under the two test modes in a foreign language testing context has seldom been reported.

Ⅲ. Research Questions

The practical and standardized nature of the SOPI is appealing in a foreign language context, since it is difficult to find a reliable pool of qualified raters and interviewers. Considering the inherent difference between the nature and frequency of opportunities to use a target language in foreign and second language contexts, it is desirable to examine whether the results from the OPI-SOPI comparisons observed in second language contexts would also hold in foreign language contexts. Acknowledging the paucity of empirical investigations on the compatibility between the OPI and the SOPI in a foreign language testing context, this study sets out to compare the performance of Korean learners of English under the OPI and the SOPI using quantitative and qualitative analyses. This study focused on the performance of advanced Korean learners of English to ensure meaningful performance samples from all participants. In sum, this study was guided by the following research questions:

(1) Are there differences in the scores of advanced Korean learners of English when tested under the two different modes of oral proficiency interview?

(2) Are there differences in the performance samples of advanced Korean learners of English when tested under the two different

modes of oral proficiency interview?

Ⅳ. Methods

1. Participants

This study recruited participants with advanced English proficiency as test takers. Since a main part of this study involved qualitative comparisons of performance samples elicited under the two test modes, it was crucial that test takers were capable of producing meaningful spoken English. The focus on advanced test takers, therefore, served as a means to ensure meaningful performance samples. Scores from the Test of English for International Communication (TOEIC) and the Test of English Proficiency developed by Seoul National University (TEPS) were employed as the criteria of general English proficiency. In particular, the TOEIC score of 900 and the TEPS score of 800 were used as the cutoff scores for the participant selection, since these correspond to the high-end of the proficiency level defined in the TOEIC and TEPS score scales.

While the TOEIC and the TEPS do not necessarily[1] include a direct measure of test takers' writing or speaking skills, they have been widely accepted in Korea as general English proficiency measures for various purposes. Practical difficulties in finding participants who took an English proficiency test measuring both receptive and productive skills precluded the use of such tests as the test taker selection criteria. In addition, high correlations between receptive and productive skills typically observed in tests measuring both receptive and productive skills (e.g., Sawaki, Stricker, & Oranje 2008) can, at least partially, justify the use of the TOEIC and TEPS scores as the selection criteria for general English proficiency.

A total of 21 undergraduate and graduate students majoring English education at a Korean university applied to participate in this study. Among these applicants, 17 were selected based on the English profi-

[1] Both the TOEIC and TEPS have Speaking and Writing modules that are optional.

ciency criteria. Two of the selected 17 had lived in English speaking countries during their childhood. These two applicants were excluded to maintain the focus of this study on a foreign language context. The remaining 15 students participated in this study as test takers. The TOEIC scores of the 15 test takers ranged from 920 to 990, and the TEPS scores from 800 to 930. All test takers studied English mainly in Korea. While there were a few test takers who studied English in English speaking countries for more than six months, none of them stayed in those countries for more than two years. The majority of the participants (12 out of 15) were female.

In addition to the test takers, an interviewer and two raters were recruited for this study. The interviewer and one of the two raters were native speakers of English. The other rater was a non-native speaker of English with extensive experiences in rating oral and written English performance samples, as well as in teaching English in Korean public schools. All three of them held bachelor's degrees in a language related major, such as English Education and Linguistics.

2. Test Instruments

This study employed two types of oral proficiency tests: a direct test (the OPI) and a semi-direct test (the SOPI). Each of the two oral proficiency tests consisted of three elicitation tasks: Description, Narration, and Expressing Opinions. The first two tasks required test takers to relate their spoken responses to given visual inputs, while the third task simply asked test takers to express their opinions on a given topic. The three tasks were chosen because of their non-interactive nature; during the response time only the test takers were supposed to produce language output. This was to control potential inconsistency of the interviewer responses in the OPI. Two closely matched sets of the three task types were selected from two preparation materials for the TSE. The key features of these tasks are summarized in Table 1 below.

Table 1. Features of Elicitation Tasks

	OPI mode			SOPI mode		
Task	Description	Narration	Expressing Opinions	Description	Narration	Expressing Opinions
Topic	a graph about eating habits	an accident at a restaurant	Ideal marriage	a graph about tourists	a quarrel at a cafe	good son and daughter
Expected Language Functions	describing, explaining	narrating, giving opinions	giving and supporting opinion	describing, explaining	narrating, giving opinions	giving and supporting opinion
Preparation time	60 sec.	60 sec.	None	60 sec.	60 sec.	None
Response time	less than 90 sec.	less than 90 sec.	less than 90 sec.	90 sec.	90 sec.	90 sec.

3. Procedures

The test takers were given one minute of preparation time for the Description and Narration tasks to interpret the given input and prepare their responses. Since there was no input material to interpret or analyze in the Opinion tasks, no preparation time was given for that task type. Response time limits were set to 90 seconds for the SOPI, while they were relatively flexible in the OPI. In particular, the interviewer was allowed to move on to the next task if a test taker finished a task in less than 90 seconds. Few test takers used more response time than 90 seconds in the OPI.

For the OPI, the interviewer was instructed to provide both verbal and non-verbal reactions to the test takers during the response time. This served two purposes: to closely simulate an actual face-to-face communication and to help the test takers feel comfortable. The interviewer was also allowed to interact with the test takers to elicit authentic and natural performance samples. However, the types of allowed interactions were limited to those pertinent to the given tasks. In addition, the interviewer was instructed to use the same manner and tone of interaction for all test takers. This was in line with O'Loughlin's (2001) design. On the other hand, interviewers in Shohamy's (1994) study were granted rather unusual amount of freedom in interacting with their test takers, while most OPI settings impose at least some degree of interviewer standardization to avoid inconsistent and potentially unfair test environments.

All tests were administered in a university phonetics lab equipped with recording devices. The OPI was administered first, followed by the SOPI. Before the test takers took the tests, instructions of the assessment procedures were provided by the author in Korean. The test takers were also told that the results of the tests, including all performance samples they would produce, would be used solely for the purpose of this study.

4. Scoring Procedures

All performance samples from the test takers under the two test modes were recorded to be scored by the two raters and transcribed for subsequent analyses. A total of 30 recordings, two test modes each consisting of the three tasks from the 15 test takers, were collected after the completion of all testing procedures. All 30 recordings were independently rated by the two raters. Prior to the rating, each rater was trained by the author to use a modified version of O'Loughlin's (2001) scoring rubric. This analytic rubric evaluated five sub-skills, which were fluency, accuracy, vocabulary, coherence and cohesion, and intelligibility. Each of these five sub-skills was scored on a six-point scale, one being the lowest and six being the highest. Each task was rated on the five sub-skills, and the sum of the five subs-skill scores within the same task was given as the task score. Since there were three tasks for each test mode, the final scores for the OPI and the SOPI were calculated as the sum of the three tasks within the same mode.

After the initial rating, score sheets from the two raters were compared. The two raters showed a satisfactory level of agreement in each sub-skill, with approximately 95 percent of the scores within one point difference. Pearson correlation coefficients between the final scores from the two raters were almost perfect, with .92 and .96 for the OPI and the SOPI, respectively. These highly correlated ratings ensured the inter-rater agreement of the both tests. Another round of rating was given for the remaining 5 percent of the initial scores for which the two raters disagreed by more than one scale. The resulting scores from the two raters were then averaged to yield the final scores of each test taker. This scoring procedure resulted in two final scores for each test taker, one for the OPI and the other for the SOPI, each of which ranged from 15 (1s on all five criteria in all three tasks by both raters) to 90 (6s on all five criteria in all three tasks by both raters).

5. Analysis of the Data

To investigate the comparability of the direct and semi-direct oral proficiency tests from multiple perspectives, both quantitative and qualitative analyses were conducted. The scores from the two tests were compared statistically and the transcripts of performance samples were analyzed using a discourse analytic approach. For the latter analysis, this study adopted Shohamy's (1994) approach, which included the comparisons of types of linguistic errors, lexical density, and other discourse features between test taker performance samples elicited from the two tests.

Ⅴ. Results

1. Score Comparison

To evaluate the comparability of the overall test taker performance under the OPI and the SOPI, the task and final scores from the two test modes were compared. In particular, the mean difference between the two sets of scores was examined using a series of paired t-tests. The results showed no significant difference between any of the mean scores, as can be seen in Table 2 below.

Table 2. Descriptive Statistics and Paired t-test Results for the OPI and SOPI scores1

	OPI Mean (S.D.)	SOPI Mean (S.D.)	*t*-statistic	p-value
Description	19.87 (4.22)	18.63 (3.66)	1.848	.086
Narration	19.23 (4.14)	18.97 (4.20)	.731	.477
Expressing Op.	17.93 (4.82)	18.20 (5.37)	-.419	.681
Total Score	57.03 (12.86)	55.79 (12.79)	1.008	.331

notes. 1. N = 15

In addition, Pearson correlation coefficients between the two sets of scores were examined. All score pairs were significantly correlated. All estimated correlation coefficients were very high. The correlation coefficient between the total scores of the two tests was .93, which is higher than Clark's (1979) .9 cutoff for using the SOPI in place of the OPI. The lowest correlation was found between the scores on the Description tasks, but even that was as high as .79. Figure 1 presents the strong linear associations between the score pairs obtained from the two test modes.

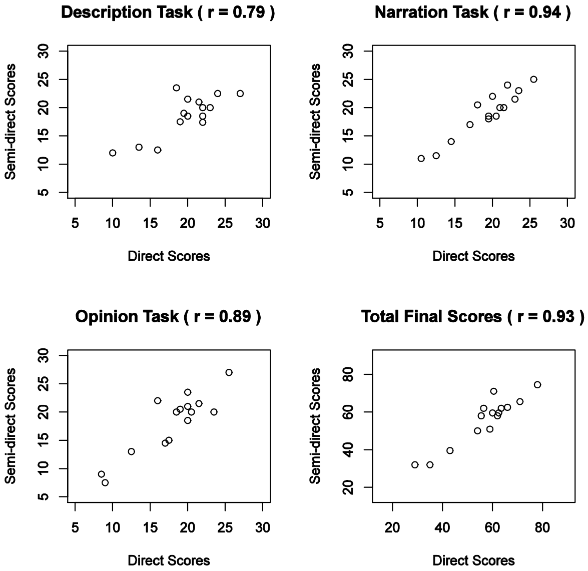

Figure 1. Test Taker Scores from the OPI and the SOPI.

2. Comparison of the Types of Linguistic Errors

In order to examine potential differences between linguistic errors elicited from the OPI and the SOPI, the types and the frequency ratios of errors elicited under the two tests were compared. The list of errors investigated in this study was developed mainly based on Shohamy (1994), but modified to appropriately capture errors that were frequently encountered in the context of this study. The resulting list included errors involving word order, verb structure, tense, gender, singular/plural, incorrect words, and subject-verb agreement.

For statistical comparisons, the frequency count of each error type was converted to the associated ratio value. That is, the count of each error type was divided by the total number of words produced by the test takers. The same ratio transformation was also used by Shohamy (1994). The resulting ratios were compared using a series of paired t-tests. The results are summarized in Table 3 below.

Table 3. Descriptive Statistics and t-test Results for the Ratio of the Error Types from the OPI and the SOPI1

	OPI Mean (S.D.)	SOPI Mean (S.D.)	*t*-statistic	p-value
Word order	.23 (.39)	.19 (.27)	.55	.591
Verb structure	.67 (.61)	1.06 (.74)	-2.35	.034
Tense	1.44 (.81)	1.77 (1.26)	-.89	.389
Gender	.02 (.08)	.17 (.29)	-1.87	.083
Singular/plural	1.05 (.74)	1.15 (.70)	-.38	.710
Incorrect words	3.99 (1.41)	3.32 (1.26)	1.71	.110
Agreement	.57 (.68)	.68 (1.04)	-.29	.776

notes. 1. N = 15

The difference between the frequency ratios of verb structure errors from the OPI and the SOPI turned out to be statistically significant at the .05 level.2) In particular, the SOPI elicited more verb structure errors than the OPI did. The performance samples from the SOPI also exhibited a slight tendency to contain more gender errors, which involved the misuse of third person singular pronouns, and fewer errors related to incorrect words than those from the OPI.

3. Comparison of the Lexical Densities

To investigate the degree of "literate-ness" of the performance samples elicited under the two tests, the lexical densities of the test taker performance from the OPI and the SOPI were compared. This comparison was motivated by Halliday's (1985) claim that more literate language samples would contain more lexical items than less literate samples would. This study adopted O'Loughlin's (2001) criteria for counting lexical items, which ignore the distinction between high and low frequency lexical items. O'Loughlin presented the median scores and the ranges of lexical densities he observed to avoid strong impacts of a few outliers. To facilitate the comparison between the results obtained from O'Loughlin's study and the results of this study, the corresponding lexical densities from each task in this study are presented in Table 4 below.

Table 4. Median Scores and Ranges of Lexical Density from the OPI and the SOPI

	OPI			SOPI		
	Description	Narration	Expressing Opinion	Description	Narration	Expressing Opinion
Median (%)	44.16	44.7	40	45.61	46.46	44
Range (%)	35.1~57.1	33.6~50.7	29.7~54.8	34.8~65.3	35.3~63.6	32.5~52.9

In O'Loughlin's study, the median scores were approximately 40 percent of lexical items,[3] while the minimum and maximum scores were ap-

[2] The Bonferroni correction of the p-values addressing the simultaneous multiple comparisons would indicate this difference as non-significant. The unadjusted p-values were reported and interpreted in this study to facilitate the comparison between the results of this study and Shohamy's (1994) results.

[3] O'Loughlin's (2001) median lexical densities for his OPI description and narration tasks were 40 percent and 38 percent, respectively. The corresponding lexical densities for the SOPI description and narration tasks were 42 percent and 41 percent, respectively.

proximately 30 percent and 50 percent of lexical items. The corresponding results from this study were very similar. This suggests that the lexical densities could be robust under different contexts and tasks.

The median lexical densities of the performance samples from the SOPI were slightly higher than those of the performance samples from the OPI.[4] This indicates that the SOPI had a tendency to elicit more literate discourses across all task types used in this study. Task types also appeared to contribute to some difference in median lexical densities. The ranking of task types in terms of the median lexical density was consistent across the two test modes. In particular, the Narration tasks elicited the highest median lexical density, followed by the Description tasks and the Opinion tasks.

4. Comparison of Discourse Features

Discourse features analyzed and reported in this section were adopted from Shohamy's (1994) analysis framework. The characteristics of each discourse feature found in the test taker performance from the OPI and the SOPI are described in the following paragraphs. Relevant excerpts from test taker utterances are provided to facilitate the interpretation. The results of the following comparisons are summarized in Table 5 at the end of this section.

(1) Rhetorical Functions and Structure

Rhetorical functions and structures of the performance samples elicited from the OPI and the SOPI turned out to be highly comparable. Since both test modes used a set of pre-determined tasks as discourse units, the sequence of "task - performance - new task - new performance" was predominant. Although a few other sequences such as "question - answer - question" occurred during the OPI, they took place only when

[4] The observed lexical densities did not differ significantly across the test modes and the task types. However, it was decided to report and interpret the median densities to address O'Loughlin's (2001) concern about outliers and facilitate the comparison between the results from this study and O'Loughlin's study.

the interviewer and a test taker encountered communication breakdowns. The following excerpts[5] from a test taker illustrate the similarity between the performance samples under the two test modes in terms of rhetorical structure and functions.

[Test taker 2: OPI, Description]

t2: uh (.) this bar graph shows how often we eat breakfast. and first of all, um- thirty eight percent of people eat breakfast everyday uh (.) which is the highest rate. and then (0.2) uh- uh- twenty two percent of people eat breakfast most days.

[Test taker 2: SOPI, Description]

t2: this chart (.) shows the countries with the most tourists uh- first of all, uh- uh (0.4) look at the highest list of the countries uh- the Fran- a France is the highest bar (0.2) in France, forty five millions of people visit there. and second, Spain um (0.2) is the second highest country- the most tourist visit.

The above excerpts show that the test taker organized her performance in the same way under the two test modes. She adopted *listing* as the main structure in both of the responses.

(2) Speech Moves

The test takers did not expand or elaborate beyond the given tasks under the SOPI. Under the OPI, however, a few instances of expansion or elaboration were observed. These occurred during the Expressing Opinion task, and often served to create a new discourse. The excerpt below provides an example of a test taker creating a new discourse by adding a personal plan to her opinion on the topic, which was 'an ideal marriage'.

[5] The numbers in the parentheses represent the duration of silence in seconds. For example, (0.2) represents approximately 0.2 seconds of silence. A dot in the parentheses represents a very short silent period.

[Test taker 5, OPI, Expressing opinion]

t5: (0.2) what I want to do in my marriage like (.) I- I'll try to have more time with my partner- my husband. and I'll try to uh (0.1) talk with him more.

The new discourse is distinguished by the use of the first person pronouns, as opposed to the exclusive use of third person pronouns in the original discourse.

(3) Communicative Properties

In general, the performance samples elicited under the OPI were slightly more interactive than the samples elicited under the SOPI. However, one-way communication was predominant under both test modes. The test takers did not attempt any explicit clarification requests or negotiation of meaning. The completion of the tasks appeared to be the sole purpose of the communication. Occasional two-way communications during the OPI took the form of a rescue from the interviewer without explicit requests from the test takers, as can be seen in the excerpt below.

[Test taker 12, Direct, Description]

```
t12: and he looks for se- seats and ( . ) uh- uh- and th- in front of
     him ( . ) um- a- a- another clerk uh (0.1) is washing- washing
     the floor uh ( . ) with long
     (0.3)
int: °mop?
t12: long- long=
int: =mop, [maybe?]
t12:       [mop], ye and um ( . ) he- he um
```

(4) Discourse Strategies

The two test modes elicited similar discourse strategies, including *repetition of parts of the question, hesitation, self-correction*, and *starting all over again*. The only difference between the performance samples from the

OPI and the SOPI in terms of discourse strategies was the lack of long silences under the OPI. This had to do with occasional feedback and/or rescue from the interviewer when she sensed long pauses.

(5) Genre, Prosodic/Paralinguistic Features, Speech Functions, and Discourse Markers

The OPI and the SOPI elicited highly comparable performance samples in terms of genre, prosodic/paralinguistic features, speech functions, and discourse markers. The genre of the performance samples elicited under the two test modes could be best categorized as either *reporting* or *monologue* in that the test takers completed the task on their own in almost all cases. The performance samples did not show any distinctive patterns or characteristics in terms of average tone and pitch shifts. Common speech functions across the two test modes included *reporting, describing,*

Table 5. Discourse Features of the OPI and the SOPI

Feature	OPI	SOPI	Comparison
Rhetorical structure/ functions	task - performance - new task (sometimes Q-A-Q)	task - performance - new task	very similar
Speech moves	description/reporting (hint of potential expansions/elaboration)	description/reporting (no sign of expansion/elaboration)	slightly different
Communicative properties	overall one-way (occasionally two-way)	one-way communication	slightly different
Discourse strategies	repetitions of parts of question, hesitation, self-correction, start all over again	repetitions of parts of question, silence, hesitation, self-correction, start all over again	slightly different
Genre	reporting/monologue	reporting/monologue	same
Prosodic/ paralinguistic features; contextualization	Laughs, hums, hesitation	Laughs, hums, hesitation, silence	very similar
Speech functions	Reporting/describing /narration	Reporting/describing /narration	same
Discourse markers	but, and, however, maybe, so, first, then, finally, even though, etc	but, and, however, maybe, so, first, then, finally, even though, etc	same

and *narrating.* These were consistently observed within the same task types regardless of the test modes. This consistency could be viewed as an indication that speech functions might be dependent on task types, not on test modes. Frequently used discourse markers included *but, and, however, maybe, so, then, finally,* and *even though,* among others.

Ⅵ. Discussion

Overall, the results from the two test modes did not differ significantly in terms of the scores and the performance samples elicited from the test takers who were advanced Korean learners of English. For the quantitative analyses of the scores, this study yielded results in line with Shohamy (1994) in that the mean score differences between the two test modes did not turn out to be significant. In addition, scores from the two test modes were all highly correlated.

As for the qualitative analysis on elicited performance samples, dramatic differences between the performance samples from the Hebrew OPI and SOPI that Shohamy (1994) found were not evident in O'Loughlin's study (2001) or in this study. The two test modes yielded essentially identical performance samples in terms of genre, speech functions, and discourse markers. Rhetorical structures and prosodic/paralinguistic features elicited from the two test modes were also highly comparable. The high comparability of the two test modes in these discourse features suggests that the global structure of the elicited performance samples did not differ across the two test modes. There were a few differences in terms of microscopic discourse features, such as speech moves and discourse strategies. Hints of potential elaboration were occasionally observed, while they seldom led to a prolonged discourse that is not directly related to the given tasks. The lack of silence and occasional two-way communications were the only instances in which the presence of the interviewer influenced the elicited performance samples.

It is noteworthy that, unlike Shohamy (1994), O'Loughlin (2001) and the present study controlled the possible task effect by matching the characteristics of elicitation tasks. In addition, the interviewers in O'Loughlin's study and this study were instructed not to interact with the test takers actively when a test taker brought up task-irrelevant topics. Therefore, Shohamy's (1994) claim that the effects of test mode overrides those of task types might be premature, since her results appear to be at least partially affected by the inherent differences in given tasks, thus leaving an influential variable uncontrolled. The fact that O'Loughlin (2001) found the evident difference between the performance samples produced under the two test modes only in a role play task, which was excluded in this study due to its interactive nature, also supports this interpretation. The impact of tasks on the performance samples was clear in this study; the elicitation tasks determined dominant speech functions.

The performance samples from the SOPI contained more verb structure errors than that from the OPI. The SOPI also elicited slightly more gender errors but fewer lexical errors. These differences in error types might be accounted for by the results of the lexical density comparison. Adopting Halliday's (1985) position, the high lexical density of the SOPI can be attributed to its tendency to elicit more literate performance samples than the OPI. Therefore, the performance samples elicited under the SOPI would be more likely to contain longer and more complex sentences, which in turn would lead to more frequent grammatical errors such as errors in verb structures. The small sample size of this study, as well as the use of multiple comparisons without adjusting for the p-values, makes the generalization of this observation difficult. However, if this finding is replicated by findings from future research, eliciting longer and more complex sentences could be considered as an advantage of the SOPI over the OPI.

Finally, the lack of clarification requests under the OPI setting is noteworthy. Unlike the participants in Shohamy (1994) and O'Loughlin (2001), the participants of this study did not, at least explicitly, ask for any help from the interviewer. This is an interesting phenomenon, and open to different interpretations. It could be attributed to task effects in that this study intentionally chose tasks that are monologic in nature, or to the advanced general English proficiency of the test takers. It is also possible that the design of this study, which paralleled the two test modes, might have given the participants an impression that the two tasks should be completed under similar conditions, although no such instruction was provided. Given the differences in terms of the frequencies of opportunities to use English for oral communication between the test takers in the two earlier studies and those in this study, the lack of experience in interactive communication using English could be another plausible explanation.

Ⅶ. Conclusions

A thorough understanding of important factors that have an impact on test taker performance is essential to develop and administer a proper oral proficiency assessment. This study made an attempt to gather empirical evidence to evaluate the comparability of the OPI and the SOPI in a foreign language testing context. Although this study focused on a small number of test takers from a narrowly defined population, the use of both quantitative and qualitative analyses allowed an in-depth investigation on the comparability of the two test modes from multiple angles. The results of this study demonstrated that the scores and the elicited performance samples from the OPI and the SOPI were in general highly comparable. This is an encouraging result, for the SOPI is more practical to administer than the OPI, especially in a foreign language context. If the results of this study could be replicated by more extensive studies based on more general learner populations, the SOPI can be safely used as a practical and more standardized way of measuring one's oral proficiency. Furthermore, the wide availability of computer technology makes it straightforward to deliver the SOPI to a number of test takers at the same time.

This study was based on a small sample of advanced English learners. While the decision to focus on advanced English learners was made based on a substantive consideration, it certainly limits the generalizability of the results. This study intentionally used tasks that were deemed monologic in nature, and therefore, the results might not generalize to tasks that are more interactive, such as role play and discussion tasks. The tasks were also given to all test takers in the same order, making the design susceptible to learning effects during the test administration. However, such learning effects were not expected to be large since the number of tasks was small and the task types were not new to the participants who studied English education. Lastly, this study did not examine individual differences between the test takers. It is possible that test takers are affected differently by the two test modes.

There remain a number of unanswered questions about the impact of test modes and task types. Future research can disentangle the impact of these important factors using controlled designs based on a larger sample size. It is essential to consider their impact under specific contexts to provide valuable insights for test takers, as well test developers and users. For example, different results are expected when inherently interactive tasks, such as role-play, are used and/or test takers with low English proficiency are asked to participate. This study investigated advanced English learners with tasks that are non-interactive by nature, and therefore evidence about the comparability between the two test modes gathered in this study should be understood and used within that specific context. Future studies can examine whether the same results would hold with larger number of test takers encompassing different English proficiency levels. Empirical evidence gathered from such extensive studies can help form a convincing argument for the use of the SOPI as a comparable yet more practical alternative to the OPI in a foreign language testing context.

References

Brown, Annie. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20.1, 1-25.

Choi, Inn-Chull. (2000). Construct validation of English simulated oral proficiency interview test method facets. *Journal of the Applied Linguistics Association of Korea* 16.1, 215-246.

Clark, John L D. (1979). Direct and semi-direct tests of speaking ability. In Eugene J. Briere & Frances B. Hinofotis, ed., *Concepts in Language Testing: Some Recent Studies* 35-49. Washington DC: TESOL.

Clark, John L D., & Swinton, Spencer S. (1980). *The Test of Spoken English as a Measure of Communicative Ability in English-medium Instructional Settings.* TOEFL Research Report, No. 7, Princeton, NJ: Educational Testing Service.

Halliday, Michael A. (1985). *Spoken and Written Language.* Melbourne: Deakin University Press.

Jeong, Hyeonjeong., Hashizume, Hiroshi., Sugiura, Motoaki., Sassa, Yuko., Yokoyama, Satoru., Shiozaki, Shuken., & Kawashima, Ryuta. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning* 61.3, 675-699.

Joo, Mi-Jin. (2007). The attitudes of students' and teachers' toward a computerized oral test and a face-to-face interview in a Korean university setting. *Journal of Language Sciences* 14.2, 171-193.

Jung, Haeng. (2000). Development of English oral proficiency test through SOPI and analysis of the results. *English Teaching* 55.2, 219-241.

Lazaraton, Anne L. (1996). Interlocutor support in oral proficiency interviews: The case of the CASE. *Language Testing* 13.2, 151-172.

O'Loughlin, Kieran J. (2001). The equivalence of direct and semi-direct speaking test. *Studies in Language Testing 13.* Cambridge: CUP.

Qian, David. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly* 6.2, 113-125.

Sawaki, Yasuyo., Stricker, Lawrence., & Oranje, Andreas. (2008). *Factor Structure of the TOEFL Internet-Based Test (iBT) : Exploration in a Field Trial Sample.* TOEFL iBT Research Report RR-04. Princeton, NJ: ETS.

Shin, Dong-Il., & Kim, Jong-Kuk. (2005). Discourse approach to face-to-face interview in an English speaking program. *The Sociolinguistic Journal of Korea* 13.2, 171-192.

Shohamy, Elana. (1994). The validity of direct versus semi-direct oral tests. *Language Testing* 11.2, 99-123.

Shohamy, Elana., Gordon, C., Kenyon, Dorris., & Stansfield, Charles. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education* 4.1, 4-9.

Shohamy, Elana., Reves, Thea., & Bejerano, Yael. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal* 40.3, 212-220.

Stansfield, Charles W. (1990, April). *A comparative analysis of simulated and direct oral proficiency interviews.* Paper presented at the Annual Meeting of the Regional Language Centre Conference, Singapore.

Stansfield, Charles., & Kenyon, Dorris. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 20.3, 347-364.

Wigglesworth, Gillian., & O'Loughlin, Kieran. (1993). An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing* 2.1, 56-67.

Ikkyu Choi
660 Rosedale Road, MS 04-R, Princeton, NJ, 08541
E-mail: ichoi001@ets.org

Received: June 28, 2014
Revised version received: July 31, 2014
Accepted: August 7, 2014