Comparability of Tasks in Assessing L2 Learners' Speaking Performance

Siwon Park

I. INTRODUCTION

Recent research efforts in task-based research have given its primary attention to identifying types of tasks that enhance learning (e.g., open-ended, structured, teacher-formed, small group, and pair work). More specifically, researchers have examined the extent to which different types of tasks impact on L2 learners' speech performance (Bygate, Skehan, & Swain, 2001; Elder, Iwashita, & McNamara, 2002; Skehan, 1996; Skehan, 1998; Skehan & Foster, 1997; Skehan & Foster, 1999; Skehan & Foster, 2001; Robinson, 1995; Robinson, 1998; Robinson, 2001a; Robinson, 2001b; Nakatsuhara, 2010; Norris, Brown, Hudson, & Yoshioka, 1998; Brown, Hudson, Norris, & Bonk, 2002; Long, 1985; Long, 1989; Long, 2005; Long & Norris, 2000; Park, 2008; Van Moere, 2010). Among these studies, a number of them examined task features and task-specific learner factors and their interactions (e.g., Skehan and Robinson's extensive research in this area), in order to better understand what makes a task more or less difficult. In addition, other researchers have been concerned about test-tasks that may help L2 educators to link teaching and assessment closely (Bygate, 2001; Douglas, 2000; Ellis, 2003; Fulcher 2003; Nakatsuhara, 2010; Nunan, 1989, 2004; Park, 2008; Skehan, 2001; Van Moere, 2010; Weir, 2005).

In language assessment, the effect of task characteristics has often been discussed in terms of task difficulty. That is, the focal interest in task-based research has been

in promoting better understanding of different cognitive characteristics of test tasks that make them either difficult or easy across different performance conditions. Motivated by different theoretical characterization of tasks, different approaches to research have been adopted and performed. Under the information-processing framework, researchers (e.g., Brown et al., 2002; Norris et al., 1998; Robinson, 1995, 2001; Skehan, 1998, 2001) have attempted to predict the cognitive demands of task characteristics on learner language production. The legitimacy of those *a priori* definitions of task characteristics that are to conjure different processing demands (i.e., different levels of task difficulty) were examined typically using interlanguage measures with the production data. In a more testing-driven conceptualization of task characteristics, difficulty of tasks was simply viewed as a method facet and was examined a posteriori using statistical methods such as Generalizability theory, Rasch analysis, and Confirmatory factor analysis.

In this study, task characteristics are interpreted and operationalized as interactant relationships and requirements in communicating information to achieve task goals and reach task outcomes. Three task types, supposedly representing different task characteristics, are examined: discussion, information gap, and one-way speech as realized in the assessment formats of the group oral discussion, information gap, and semi-direct speaking tasks respectively. They are distinct at the structural level, and their task features as described in Table 1.

Table 1

Task types: Interactant (X & Y) Relationships and Requirements in Communicating Information to Achieve Task Goals and Reach Task Outcomes (adapted from Pica, Kanagy, and Falodun, 1993)

		Group oral tasks		Semi-direct (e.g., picture description)
		Topic discussion	Information gap (Jigsaw)	
1	Information Holder	X = Y	X or Y	X
2	Information Requester	X = Y	Y or X	None
3	Information Supplier	X = Y	X or Y	X
4	Information Requester-supplier relationship	2 way>1 way (X to Y & Y to X)	2 way (X to Y/Y to X)	None
5	Interaction requirement	-required	+required	-required
6	Goal orientation	-convergent	+convergent	-convergent
7	Outcome options	1+/-	1	Not specified

Considering the seven characteristics across the three tasks in Table 1, unique characteristics of each are noticeable. Thus, the three test tasks offer task-specific aspects that may help elicit different language output from the candidates in this study.

Thus, the purpose of the current study is to examine the extent to which three second language (L2) speaking test tasks impact on L2 examinees' speech production: topic discussion, information gap, and semi-direct speaking test tasks. In order to achieve the purpose, the research question is set as "How comparable are the three sets of speech samples produced by examinees in performing the three test-tasks?"

II. METHOD

1. Participants

A total of one hundred twenty two Japanese learners of English who are English majors at a university in Japan participated as the examinees in this study. The participants vary in their class standing: twenty-one juniors, sixty-five sophomores, and thirty-three freshmen. Out of one hundred twenty-two, forty-one students were male (34%) and the rest female. Information about their English proficiency measured by an in-house English proficiency exam (named KEPT) which was administered seven months prior to the study for the seniors, juniors, and sophomores and five months ago for the freshmen is shown in Table 2 below. Even though there is a considerable time lag between the data collection and the KEPT administrations, the KEPT scores of 87 examinees are presented in Table 2 to demonstrate variation in the examinees' overall proficiency levels. Not all examinees took the KEPT; therefore, the scores come from only 87 examinees (out of 122). As the table indicates, there is indeed much variation among the examinees in terms of their English proficiency and especially in their speaking ability.

Table 2

Examinees' English Proficiency Measured by KEPT (N = 87)

	M	*SD*	Range	Min	Max
Total	64.96	10.18	47.87	47.33	95.20 (/100)
Speaking only	12.60	2.83	12.75	6.93	19.68 (/20)

Note. Total is a composite score of reading, listening, grammar, writing and speaking section scores of the test.

In addition to learner participants, ten raters who have experiences in assessing L2 learners' speaking participated in this study. The raters were selected from the pool of full-time instructors at the English Language Institute of the university where the examinees were enrolled. They all had previously served as a rater for the speaking section of the KEPT in January and March of the year when the testing experiment was conducted. They are all native speakers of English holding a master's degree in applied linguistics or related fields. Out of ten, three raters were female, and the rest male. Five raters were from the U.S., two from the U.K., two from Australia, and one from Singapore. All of them reported that they had experience teaching Japanese EFL learners at a university prior to where this study was conducted. Their teaching period at the universities varied from three years to thirteen years.

2. Test Instruments

As mentioned earlier, group oral discussion, information gap, and semi-direct speaking tasks were examined in this study. The three tasks were chosen because they are the most common task types that have been used in the L2 classroom and assessment. Hence, there is no structural manipulation to elicit specific speech samples from the examinees unlike those theory-driven task-based studies that have previously been conducted by Skehan, Robinson and their colleagues. Rather, the tasks in this study are known to have a *priori* characteristics that are fundamentally different from each other. For more information about these tasks, refer to Park (2008).

3. Procedure

Testing experiments were conducted over a few days. The sessions for the first two days were to collect the testing and survey data, and the last day was mainly to

conduct the stimulated think-aloud protocols (the result of which is not reported in the paper).

For the group oral testing, a group of four students sat together for one session. The presentation order of the tasks was counter-balanced and so was raters' rating order of the tasks. There were some groups of examinees that stayed in the same test room and were given the tasks to be performed without changing the group members. There were other groups of examinees who would have to change their test room in order to take the test with different examinees. This procedure was implemented to examinee if the familiarity between/among group examinees affected their performance in group oral.

As for the administration of the semi-direct speaking test, three versions of the semi-direct speaking test that contained the same tasks with differing orders were prepared for this study. The first part of the test included a Japanese instruction and five warm-up questions. The test was delivered via computer so that test materials were displayed with a timer while the sound was being played. Next to the computer monitor, a video-camera was set up so that the speech produced by an examinee could be recorded clearly.

In each session of group oral, two raters scored the same examinees interacting in a group. Once a group of four examinees all sat in the seats labeled from A to D for the identification purpose, one of the raters told the examinees that they had two minutes of planning time for discussion. Once two minutes had elapsed, a rater once again told the examinees that they had 12 minutes for their discussion on the topic written on the prompt or for completing the task described on the task prompts (in case of information gap). Time for planning and task completion was strictly monitored by the raters using a stopwatch so that the amount of time spent for task completion could be controlled constant across all testing sessions.

In the case of group oral, ratings were completed at the testing site, i.e., rating was done concurrently. However, ratings for the semi-direct test were completed using video recordings of examinees' responses to the test tasks, once all sessions were completed. Video-recordings of the responses were rendered to a computer from the video-cameras, and several samples as a set were saved onto a CD so that those samples could be viewed using Windows Media Player. Raters were assigned 8-10 randomly selected CDs for their individual ratings. All samples were double-rated with the responses from the semi-direct speaking tasks.

4. Descriptive Statistics and Data Screening

Before analyzing the data for the analyses of variance (ANOVA), the distributions of the data were examined to check that the characteristics of the data meet the assumptions of parametric statistical tests. In examining the normality of the distribution of the data, information from three types of techniques was utilized – Z-transformed skewness and kurtosis statistics, Kolmogorov-Smirnov Normality Test, and boxplots of non-normal data. Therefore, decisions regarding the normality of the data distribution were made based on the information collected from these three techniques. As the ANOVA procedure is known to be robust against marginally non-normal data, a loose criterion was applied to rejecting the normality assumption. When a data set was detected violating the normality condition unanimously by the three techniques, an alternative analysis method, e.g., a non-parametric test, was deemed appropriate to deal with its non-normality.

Table 3 and Figure 1 report the descriptive statistics of the Rasch adjusted scores obtained from examinee performance on the three tasks. In case of the semi-direct speaking test, as there were three sub-tasks, Rasch adjusted scores resulting from the FACETS analysis were used to calculate the descriptive statistics. Descriptive

statistics of the three semi-direct speaking tasks are reported in Table 4.

Regarding the data screening, in order to maintain a complete set of data without missing/incomplete responses, the listwise deletion method was applied. In total, 94 score sets survived from this deletion method and were subjected to the analysis. Loss of the data (i.e., scores from 28 examinees) occurred mainly because the examinees did not participate in one or more testing sessions. In some cases, examinees' performance on Task 3 was not recorded properly, and all responses of those examinees including ones collected using other tasks were deleted from the final data set.

The mean scores of the categories under the Topic discussion and Information gap tasks are all close to 6.0. Also, their standard deviations (SD) are similar to each other. However, the range values reveal that there was more variation in score assignment with the Information gap task, although the combined scores for the categories of the semi-direct speaking test show slightly different patterns in score distribution. Except for the Task-completion category, means for other categories are generally low. Also, there is a notable difference in the mean values between the Task-completion category and other categories of the semi-direct speaking test. In addition, the Task-completion category is the only one in which examinees were assigned the full score of 9.0.

As mentioned earlier, for the normality check of the data sets, several techniques were performed. First, the Z-transformed skewness or kurtosis values reveal that there were six categories with non-normal data. Those values are underlined in Table 3. Among those six categories, the Kolmogorov-Smirnov Normality test identified four categories as being non-normal, but the test newly identified the Jigsaw-Grammar data as being non-normal as well. Finally, boxplots were examined of those data sets identified as non-normal, and they are shown in Figure 1.

Table 3

Summary Statistics of the Scored Responses for Topic Discussion, Information Gap, and Semi-direct Speaking Tasks (N = 94)

	Topic discussion				
	Pron	Flue	Gramm	Vocab	Interact
M	6.31	6.34	6.05	6.13	6.60
SD	0.95	1.16	0.93	1.02	1.22
Range	4.88	5.82	4.81	4.39	5.05
Min	4.00	3.00	4.00	4.39	3.85
Max	8.88	8.82	8.81	8.78	8.90
$Z_{Skewness}$	2.03*	0.15	2.69*	2.25*	-0.01
$Z_{Kurtosis}$	0.11	-0.17	0.61	-0.78	-1.47
	Information gap				
	Pron	Flue	Gramm	Vocab	Interact
M	6.28	6.47	6.07	6.02	6.86
SD	1.03	1.10	0.93	0.96	1.23
Range	5.33	5.07	4.72	5.21	5.64
Min	3.57	3.77	4.08	3.71	3.26
Max	8.90	8.84	8.80	8.92	8.90
$Z_{Skewness}$	-0.36	-0.17	1.69*	2.50	-2.27
$Z_{Kurtosis}$	0.22	-0.64	-0.33	1.30	0.67
	Semi-direct speaking (combined)				
	Pron	Flue	Gramm	Vocab	Interact
M	5.93	5.92	5.65	5.73	7.85
SD	1.24	1.20	0.86	0.96	1.06
Range	6.88	6.17	3.86	5.17	5.05
Min	2.00	2.67	3.67	2.98	3.95
Max	8.88	8.84	7.53	8.15	9.00
$Z_{Skewness}$	-0.06	-0.74	0.29	-1.07	-6.29*
$Z_{Kurtosis}$	0.27	0.02	-0.77	1.00	5.84

Note. Pron = Pronunciation, Flue = Fluency, Gramm = Grammar, Vocab = Vocabulary, Interact = Interaction, Compl = Task-completion.
The asterisk (*) indicates a significant result of the K-S test on the particular category data.

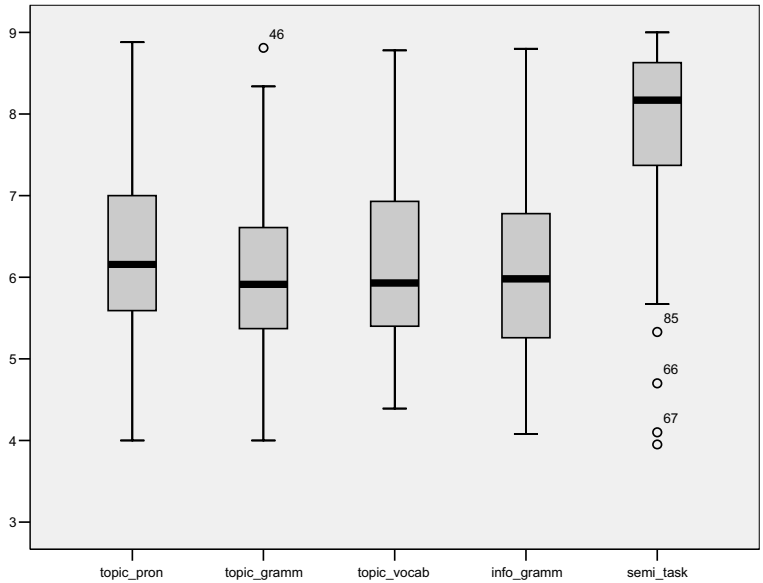

Figure 1. Boxplots of the variables which resulted as non-normal in data distribution
Note. Circles indicate outliers. topic_pron = Topic pronunciation, topic_gramm = Topic grammar, topic_vocab = Topic vocabulary, info_gramm = Information grammar, semi_task = Semi-direct speaking task-completion.

Among the five boxplots, the one for the Task-completion of the semi-direct speaking test was found to have outlying cases. Also, the distribution was not nearly symmetrical, i.e., it was significantly negatively skewed. A closer look into the data from the whole semi-direct speaking test appears necessary in order to identify the source for the significant non-normality of the category, and Table 4 and Figure 2 are presented next for that purpose.

The same notation scheme is used – underlined values indicate problematic items based on their $Z_{skewness}$ or $Z_{kurtosis}$ values, and the asterisks are to flag a

significant result of the Kolmogorov-Smirnov Normality Test. Several categories are identified as having with distributional problems. All of the category data except for Pronunciation and Fluency were identified as problematic in their distributions.

Considering the information regarding the normality of the data all together, a decision was made to exclude five data sets in the analysis – two data sets for Interaction of Topic discussion and Information gap and three data sets for Task-completion of the three semi-direct speaking tasks. Data transformation (e.g., using the reciprocal transformation ($1/X_i$)) was not deemed desirable as the degree of skewness and/or kurtosis was too significant for the approach to produce any sufficient distributional changes. In addition, interpretation of results based on the transformed data may be difficult especially in relation to other variables. There were other data sets found to be marginally non-normal, but the degree was not serious and relatively inconsequential considering the robustness of the ANOVA procedure.

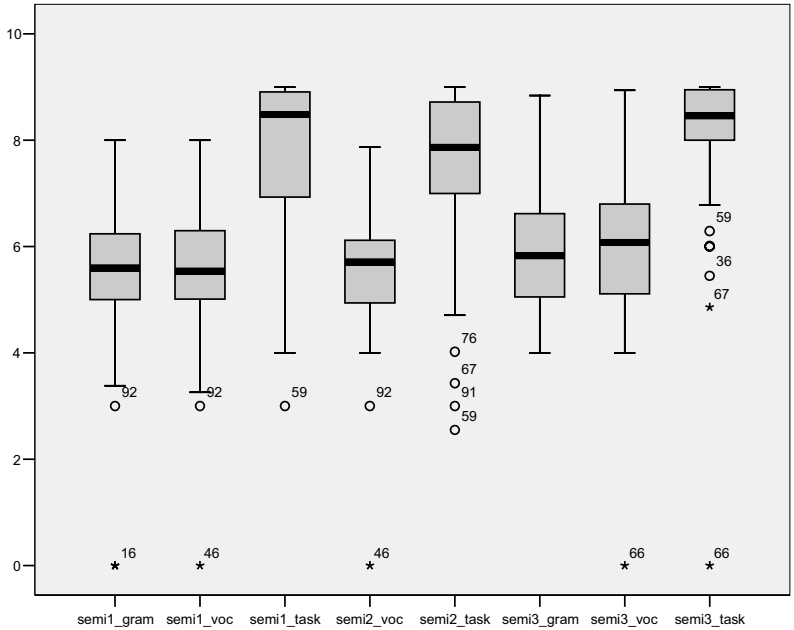

Figure 2. Boxplots of the variables which resulted as non-normal in data distribution
Note. Circle indicates outliers and star extreme cases. semi1_gram = Semi-direct speaking grammar,
semi1_task = Semi-direct speaking task-completion, semi2_voc = Semi-direct speaking vocabulary,
semi2_task = Semi-direct speaking task-completion, semi3_gram = Semi-direct speaking grammar, semi3_voc = Semi-direct speaking vocabulary,
semi3_task = Semi-direct speaking task-completion.

Table 4

Summary Statistics of the Scored Responses for the Three Semi-direct Speaking Tasks (N = 94)

Semi-direct speaking					
	Semi-direct 1 (Picture task)				
	Pron	Flue	Gramm	Vocab	Compl
M	5.92	5.92	5.48	5.61	7.82
SD	1.32	1.29	1.27	1.17	1.37
Range	6.93	6.87	8.00	8.00	6.00
Min	2.00	2.00	0.00	0.00	3.00
Max	8.93	8.87	8.00	8.00	9.00
$Z_{Skewness}$	-0.14	-0.87	-6.53*	-4.31*	-5.36*
$Z_{Kurtosis}$	-0.25	0.66	11.96	9.74	2.82
	Semi-direct 2 (Map task)				
	Pron	Flue	Gramm	Vocab	Compl
M	5.83	5.80	5.55	5.52	7.54
SD	1.28	1.24	1.01	1.05	1.39
Range	6.83	5.88	5.20	7.87	6.45
Min	2.00	3.00	3.65	0.00	2.55
Max	8.83	8.88	8.85	7.87	9.00
$Z_{Skewness}$	0.30	-0.42	2.00	-6.37*	-5.87*
$Z_{Kurtosis}$	-0.18	-0.58	1.26	14.63	4.81
	Semi-direct 3 (Speech task)				
	Pron	Flue	Gramm	Vocab	Compl
M	6.05	6.06	5.94	6.05	8.18
SD	1.26	1.31	1.06	1.26	1.22
Range	6.87	5.83	4.84	8.94	9.00
Min	2.00	3.00	4.00	0.00	0.00
Max	8.87	8.83	8.84	8.94	9.00
$Z_{Skewness}$	-0.35	-1.04	1.98*	-3.89*	-15.47*
$Z_{Kurtosis}$	0.12	-0.75	-0.29	9.42	43.57

Note. Pron = Pronunciation, Flue = Fluency, Gramm = Grammar, Vocab = Vocabulary, Compl = Task-completion.
The asterisk (*) indicates a significant result of the K-S test on the particular category data.

III. RESULTS

1. Repeated-Measures ANOVAs with the Three Test Tasks

A repeated measures ANOVA was performed on the three effects, i.e., three test tasks (Topic discussion, Information gap, and Semi-direct speaking), four *categories* under each test task (Pronunciation, Fluency, Grammar, and Vocabulary), and their interaction (*test by category*). First, descriptive statistics in Table 5 are presented only with mean and standard deviation values once again to help interpret the results in this section.

Table 5

Descriptive Statistics of the Scored Responses for the Three Test Tasks (N = 94)

Tasks	Categories	Mean	SD
Topic discussion	Pronunciation	6.31	0.95
	Fluency	6.34	1.16
	Grammar	6.05	0.93
	Vocabulary	6.13	1.02
Information gap	Pronunciation	6.28	1.03
	Fluency	6.47	1.10
	Grammar	6.07	0.93
	Vocabulary	6.02	0.96
Semi-direct speaking	Pronunciation	5.93	1.24
	Fluency	5.92	1.20
	Grammar	5.65	0.86
	Vocabulary	5.73	0.96

Table 6 reports the result of the ANOVA. First, Mauchly's Tests of Sphericity with the two main effects and one interaction in the model indicate that the assumptions of sphericity are violated with all the effects – with the main effect of test, $\chi^2(2) = 19.26$, $p < 0.01$, with the main effect of *category*, $\chi^2(5) = 29.91$, $p < 0.01$, and with the interaction effect of *test by category*, $\chi^2(2) = 64.56$, $p < 0.01$. Therefore degrees of freedom have been corrected using the Greenhouse-Geisser estimates of sphericity.

Table 6

Analysis of Variance for the Effects of Test Tasks, Categories, and Their Interaction

Source	*SS*	*df*	*MS*	*F*	η^2
Test	39.58	1.68	23.53	18.14*	0.16
Error(test)	202.89	156.45	1.30		
Category	21.16	2.57	8.23	19.76*	0.18
Error(category)	99.58	239.18	0.42		
Test by Category	2.15	4.83	0.44	1.47	0.02
Error(test by category)	136.16	449.61	0.30		
Total	501.52	854.33			

Note. * $p < 0.05$.

All effects are reported as significant at $p < 0.01$. There was a significant main effect of the *test* tasks, $F(1.68, 156.49) = 18.14$, $p < 0.01$. *Post hoc* comparisons in Table 7 revealed that rating scores of Topic discussion and Information gap tasks were significantly higher than that of the semi-direct speaking test. But no significant difference was found between the two group oral tasks. In addition, the other main effect, *category* was also found significant, $F(2.57, 239.18) = 19.76$, $p < 0.01$. The *post hoc* comparisons in Table 8 identified that the significant difference(s) between the categories reside between (a) Pronunciation and Grammar, and (b) Pronunciation

and Vocabulary, as examinees receiving higher ratings on Pronunciation than Grammar or Vocabulary. In addition, examinees received significantly higher ratings on Fluency than Grammar or Vocabulary.

Table 7

Multiple Comparisons for the Main Effect of Test

(I) test	(J) test	Mean Difference (I-J)	Std. Error
1	2	0.00	0.06
	3	0.40*	0.08
2	1	0.00	0.06
	3	0.40*	0.09
3	1	-0.40*	0.08
	2	-0.40*	0.09

Note. * $p < 0.05$, 1 = Topic discussion, 2 = Information gap, 3 = Semi-direct speaking.
α adjusted for multiple comparisons by the number of comparisons (α/3).

Table 8

Multiple Comparisons for the Main Effect of Category

(I) category	(J) category	Mean Difference (I-J)	Std. Error
1	2	-0.07	0.05
	3	0.25*	0.05
	4	0.21*	0.06
2	1	0.07	0.05
	3	0.32*	0.05
	4	0.29*	0.06
3	1	-0.25*	0.05
	2	-0.32*	0.05
	4	-0.04	0.03
4	1	-0.21*	0.06
	2	-0.29*	0.06
	3	0.04	0.03

Note. * $p < 0.05$, 1 = Pronunciation, 2 = Fluency, 3 = Grammar, 4 = Vocabulary.
α adjusted for multiple comparisons by the number of comparisons (α/6).

Finally, for the interaction effect of *test by category,* the analysis revealed that there was no significant interaction and that the effects of the categories did not significantly differ across the three test tasks. Looking at the interaction in Figure 3, ratings assigned to examinees were higher on all four categories of the two tasks – Topic discussion and Information gap – than those of the semi-direct speaking test. On all categories, rating scores were dropped for the semi-direct speaking test, although ratings were relatively similar for the examinees' performance on Topic discussion and Information gap tasks.

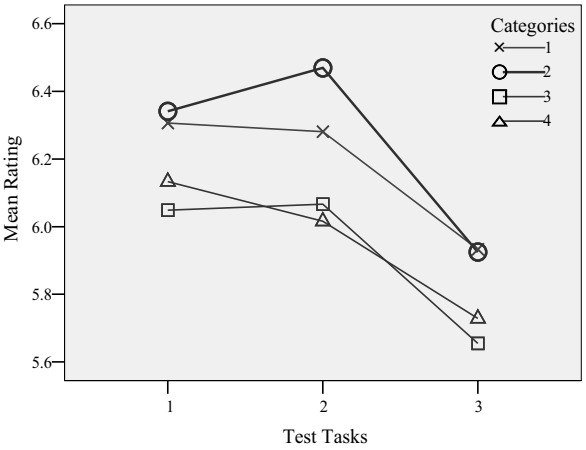

Figure 3. Rating score plots of the three tasks.
Note. 1 = Pronunciation, 2 = Fluency, 3 = Grammar, 4 = Vocabulary.

2. Univariate Tests with Each Category Measure of Three Test Tasks

A question that has not been answered yet is one regarding the contrasts of each category measure across three test tasks. Figure 3 above shows the pattern of mean ratings of the four categories across three test tasks. Yet, it has not been confirmed

whether or not the difference among the four categories within one test task is statistically meaningful. In order to address the question, univariate tests with each category measure across three test tasks were performed, and the result is reported in Table 9.

Table 9

Univariate Analysis for Individual Categories across the Test Tasks

Category Measures	*SS*	*df*	*MS*	*F*	Sig.	η^2	Power
Pronunciation	8.16	1.56	5.25	7.87	0.002	0.08	0.90
Fluency	15.24	1.80	8.47	16.39	0.000	0.15	1.00
Grammar	10.17	1.81	5.60	13.71	0.000	0.13	1.00
Vocabulary	8.16	1.79	4.56	8.71	0.000	0.09	0.95

First, Mauchly's Tests of Sphericity with the four measures all indicated that the assumptions of sphericity are violated – with Pronunciation, $\chi^2(2) = 30.96$, $p < 0.01$, with Fluency, $\chi^2(2) = 10.83$, $p < 0.01$, with Grammar, $\chi^2(2) = 9.93$, $p < 0.01$, and with Vocabulary, $\chi^2(2) = 11.63$, $p < 0.01$. Therefore, considering and reporting the effects, degrees of freedom have been corrected using the Greenhouse-Geisser estimates of sphericity, and the subsequent interpretation will be based on the corrected *F*-values.

When reading the results in Table 9, a caution needs to be taken with regards to the significance level. Since all the measures are under the same high-order trait (speaking ability), and the categories are known to be highly correlated, interpretation of one outcome is more or less dependent on that of the others. However, as noticeable from the values for the significance, they all came out low; hence, it was decided not to consider any subsequent adjustment of significance levels with the measures.

As noted in Table 9, all category measures resulted in meaningful differences across three measures for the same category for the three tasks, i.e., $p \leq .002$. In

order to understand where the significance is reflected, *post hoc* comparisons were performed, and the result is reported in Table 10.

Table 10

Multiple Comparisons of the Categories across the Three Test Tasks

Measure	(I) factor1	(J) factor1	Mean Difference (I-J)	S.E.
Pronunciation	1	2	0.025	0.075
	1	3	0.373*	0.109
	2	3	0.348*	0.125
Fluency	1	2	-0.128	0.082
	1	3	0.416*	0.103
	2	3	0.545*	0.111
Grammar	1	2	-0.018	0.077
	1	3	0.394*	0.086
	2	3	0.411*	0.101
Vocabulary	1	2	0.117	0.081
	1	3	0.405*	0.110
	2	3	0.288*	0.106

Note. * $p < 0.05$, 1 = Topic discussion, 2 = Information gap, 3 = Semi-direct speaking. α adjusted for each of multiple comparisons by the number of comparisons (α/3).

Table 10 shows that the significance in the main effects is due to the differences between the measures of Topic discussion or Information and those of Semi-direct speaking test. There was no meaningful difference found between Topic discussion and Information gap tasks with their category measures.

IV. DISCUSSION AND CONCLUSION

The purpose of this study was to examine the comparability of the three test tasks reflected in performance scores, that is, how comparable the three tasks are in assessing L2 learners' speaking ability reflected in the individual and combined scores of the five categories described in the rating scale. In order to address the question, the ability scores were examined using analysis of variance (ANOVA).

The ANOVA procedure revealed that the tasks could not be considered comparable in difficulty. The ability scores adjusted using Rasch analyses were examined for the comparability across the tasks. The ANOVA produced a significant main effect of tasks at $F(1.68, 156.49) = 18.14$, $p < 0.01$. A further examination with *post hoc* comparisons indicated that both the score means of Topic discussion and of the Information gap tasks were significantly different from that of semi-direct speaking test. However, the two group oral tasks were not statistically different. The two groups of tasks and three semi-direct speaking tasks are not comparable in difficulty.

The findings in this study suggest that the tasks examined in this study may be different in the degree of task demands so that the different tasks required the examinees to apply differing degrees of cognitive and linguistic abilities in responding to them. In measurement, this would mean that one cannot draw the same type of inference about examinees' ability based on their performance on different tasks. That is, the inference about one examinee's ability of L2 oral proficiency on one task wound not be comparable to another inference drawn from his/her performance on another task.

REFERENCES

Brown, J., D., Hudson, T., Norris, J., & Bonk, W. (2002). *An investigation of second language task-based performance assessments.* (Technical Report #24). Honolulu: University of Hawai`i, Second Language Teaching & Curriculum Center.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 23-48). Harlow, UK: Pearson Educational Limited.

Bygate, M., Skehan, P., &Swain, M. (2001). *Researching pedagogic tasks: second language learning, teaching and testing.* Harlow, UK: Pearson Educational Limited.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: CUP.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347- 368.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

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

Long, M. (1985). A role for instruction in second language acquisition: task-based language teaching. In K. Hyltenstam and M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (Vol. 18, pp. 77-99). Clevedon, Avon: Multilingual Matters Ltd.

Long, M. (1989). Task, group, and task-group interactions. *University of Hawai`i Working Papers in ESL, 8*(2). Honolulu: University of Hawai`i, Department of English as a Second Language.

Long, M. H. (2005). *Second language needs analysis.* Cambridge: Cambridge

University Press.

Long, M. H., & Norris, J. M. (2000). Task-based teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597-603). London: Routledge.

Nakatsuhara, F. (2010). *Interactional competence measured in group oral tests: how do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups?* Paper presented at the Language Testing Research Colloquium, April, 2010, Cambridge.

Norris, J., Brown, J., D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* (Technical Report #18). Honolulu: University of Hawai`i, Second Language Teaching & Curriculum Center.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Nunan, D. (2004). *Task-based language teaching.* Cambridge: Cambridge University Press.

Park, S. (2008). *An exploration of examinee abilities, rater performance, and task differences using diverse analytic techniques.* Unpublished doctoral dissertation, University of Hawaii, Honolulu, HI.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communicative tasks for second language instruction. In G. Crookes and S. M. Gass (Eds.), *Tasks and language learning: integrating theory and practice* (pp. 9-34). Clevedon: Multilingual Matters.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning,* 45, 99-140.

Robinson, P. (1998). State of the art: SLA theory and second language syllabus design. *The Language Teacher, 22*(4), 7-14.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics,* 22(1), 27-57.

Robinson, P. (2001b). Task complexity, cognitive resources, and syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and Second language instruction* (pp. 287-318). Cambridge: Cambridge University Press.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17,* 38-62.

Skehan, P. (1998). *A cognitive approach to language learning.* New York: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 167-185). Harlow, UK: Pearson Educational Limited.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influence on foreign language performance. *Language Teaching Research, 1*(3), 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49,* 93-120.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). New York: Cambridge University Press.

Van Moere, A. (2010). *Group oral tests: What kinds of tasks and functions are optimal for eliciting and measuring interactional competence?* Paper presented at the Language

Weir, C. J. (2005). *Language testing and validation: an evidence-based approach.* New York: Palgrave Mcamillan.