Learners' perception of task difficulty and actual performance in L2 speaking

Siwon Park and Megumi Sugita

I. INTRODUCTION

A growing body of research has revealed that a task may not be a neutral device for assessing L2 speaking (Brindley, 2000; Brindley & Slayter, 2002; Bygate, 1999; Wigglesworth, 2001). Different tasks demand the operation of different language skills from L2 learners; therefore, an inference of an examinee's L2 speaking ability drawn from his/her performance only on a single task may be entirely invalid or partially valid at best. Such possibility begs for more empirical research on the differential effects of tasks on examinee performance through cross-examinations of various task types. That is, more research effort is needed to better understand the systematic effects of differing tasks on examinees' speech performance. In addition, more research findings are required to promote further understanding of to what extent factors such as the task characteristics, examinees' perceptions of the tasks, or rater characteristics affect examinees' performances on the tasks (Brown, 2003).

Among the factors mentioned above as known to affect examinee performance, the current study concerns examinees' perceptions of the tasks, i.e., if and to what extent their perceptions of the tasks are related to their performance. By employing three oral tasks – topic discussion, information gap, and semi-direct speaking, this study attempts to examine the extent to which examinees' perception concurs with their actual performance on the tasks, or vice versa. Correlations are examined between survey responses and the rating scores to see if there are any meaningful relationships between the perception and performance aspects.

1. Task difficulty

What makes a given task more or less difficult for a learner is of great concern among L2 researchers and teachers involved in task-based teaching and syllabus design. In an attempt to conceptualize task difficulty, researchers have identified various factors that appear to influence task performance, such as "input," "procedures," and "the learner" (Nunan, 2004), "cognitive load," "communicative stress," "particularity and generalizability," "code complexity and interpretive density," "process continuity" (Candlin, 1987), "cognitively defined task complexity," "learner perceptions of task difficulty," and "interactive conditions under which tasks are performed" (Robinson, 2001). Robinson (2001) further distinguishes "complexity" from "difficulty," in that "complexity" concerns a feature of the task whereas "difficulty" is operationalized in terms of perceptions of task difficulty on the part of learners.

While there are a number of studies that looked into different facets of task difficulty (Robinson, 2001; Skehan, 2001; Skehan & Foster, 2001), not many of them examined the relation between the perceptional aspect of task difficulty and its relationship to task performance (Elder, Iwashita, & McNamara, 2002). A general consensus as to the relationship between the two factors, however, is that task difficulty cannot be reliably demonstrated as an estimate in post hoc manner (Elder et al., 2002).

2. Examinees' perceptions in task difficulty

The differential effects of tasks need to be understood from multiple aspects including test method, cognitive skills, and *psychological attachment* with the performance conditions involved, in addition to the ability trait to be measured. For instance, there could be tasks that require skills more cognitively invariant. Or, there may be cases where the simplest task (e.g., discussion) in format requires a more cognitively charged process for an examinee to perform the task. Snow (1993) argues that there may be tasks

that require more psychological involvement as opposed to a rather neutral one; hence, examinees may approach them with a fairly stable manner. Although this psychological aspect is somewhat difficult to conceptualize compared to others discussed earlier, it is crucial to understand such aspect as it will differentially affect examinee performance patterns, i.e., they may lead the examinee to activate and apply different strategies in responding to given tasks.

In order to examine the aspect of examinees' psychological involvement in the three tasks in this study, a survey instrument was developed and administered. The items included in the survey were mainly designed to inquire how examinees considered the relative difficulty of the three tasks and which task types were perceived most relevant to the testing of their English speaking skills. Their perception of the task difficulty and the actual performance on the tasks were compared, and the results are reported in the subsequent sections.

II. METHOD

1. Participants

The survey and test data come from a total of 87 Japanese learners of English who are English majors at a university in Japan. Out of 87, twenty eight students were male and the rest female. Information about their English proficiency measured by an English speaking test which was administered approximately six months prior to the study is shown in Table 1 below. As the table indicates, there is much variation among the examinees in terms of their English speaking proficiency.

Table 1 ***Examinees' English Proficiency (N = 87)***

	M	*SD*	Range	Min	Max
Speaking	12.60	2.83	12.75	6.93	19.68 (/20)

2. Test instruments

The three tasks employed in this study are neither highly specified nor highly structured. They 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. Rather, the tasks in this study are known to have *a priori* characteristics that are fundamentally different from each other. Table 2 is presented below, in order to highlight seven differential characteristics of the three task types.

Table 2 ***Task Types: Interactant (X & Y) Relationships and Requirements in Communicating Information (INF) to Achieve Task Goals and Reach Task Outcomes (adapted from Pica et al., 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

3. Research design

Table 3 summarizes the design of the research. As presented in Table 2, there are three main tasks, two of them in group oral and one in individual speaking formats. The group oral tasks involved three to four students depending on their availability. For Task 1, examinees performed topic discussion in group oral and for Task 2, a two-way jigsaw task also in group oral. For Task 3, examinees performed three tasks in semi-direct speaking format. The three tasks were picture, map, and speech tasks. For the first two tasks, rating was done concurrently while examinees were performing the given tasks, while the rating for Task 3 was done using recordings of the examinee performance on the tasks. All examinee responses were double-rated, i.e., by two raters. Regarding the overall rating design, all raters/all examinees design was adopted for easier and more complete data analyses in this study. Therefore, examinees were rated by all raters, and all examinees took all three tasks. The design helped accommodate a full sub-data connection for the MFRM analyses. Examinees were asked to complete the follow-up survey as soon as they finished the last testing session. The major part of current study concerns the survey responses analyzed together with the rating scores.

Table 3 ***Summary of the Research Design (N=78)***

	Task 1	Task 2	Task 3
Arrangement	Group oral	Group oral	Individual
# of examinee	3 ~ 4	3 ~ 4	1
Task type	Topic discussion	Information gap (jigsaw)	Semi-direct
# of sub-tasks	N/A	N/A	3 (picture, map, & speech)
Rating	Done concurrently; Double-rating	Done concurrently; Double-rating	Done using recordings; Double-rating

	Task 1	Task 2	Task 3
# of raters	2	2	2
Design	All raters; all examinees	All raters; all examinees	All raters; all examinees

III. ANALYSIS

In order to examine if there were any meaningful relationships between examinees' performance on the test tasks and their perception of their relative difficulty, a series of bivaraite correlational analyses (using Pearson coefficients) were calculated. A survey instrument was prepared to examine the relationship, and examinees were asked to complete it after finishing all three test sessions. The survey consisted of questions regarding perceptions on their own English proficiency, difficulty of the tasks, levels of their performance on the tasks, and the like. Among the questions, ones that appeared most relevant to the research purpose of the current study were analyzed using Pearson correlations.

1. Data screening

Among the responses from all 78 examinees, those with relatively many missing items were eliminated. After the listwise deletion procedure was applied, responses from 66 examinees remained and were subjected to the analyses together with their rating scores on the test tasks.

2. Relationship between perception and performance on the tasks

The first analysis with the survey data concerns if there is any meaningful relation between examinees' perception of their English ability and their actual speaking performance. English ability is expressed and measured on five speaking skill-components using the rating scale (i.e., Pronunciation, Fluency, Grammar, Vocabulary, and Interaction/Task-

completion). Table 4 reports the descriptive statistics of the items included in the analysis. It includes both examinees' test scores and their responses to five survey questions, concerning the perception on their English skill levels and also on the difficulty of the three main test tasks that they performed for the study.

Mean and standard deviation (*SD*) values were similar across the five test scores except for the one of Interaction. Among the three test tasks, examinees received the highest rating on the information gap task and the lowest on the semi-direct speaking test. As for the mean and *SD* values for the survey responses, values are similar across the five skills items, although the mean for Interaction is lower than the other scores, a reversed pattern of the test scores. The perceived ability were measured on a 5-point Likert scale (from "very good" – to "very poor"), and the difficulty perception on a 6-point scale (1-6, 1 being the easiest and 6 being the most difficult).

Among the four linguistic measures, examinees performed best in Fluency followed by Pronunciation, but performed worst in Grammar. On the contrary, they perceived Grammar as their strongest skill, followed by Vocabulary; yet, Pronunciation being the weakest. Regarding the difficulty of the tasks, examinees considered the semi-direct speaking test the most demanding while regarding the information gap task as the easiest. On their actual performance side, however, they performed worst on the semi-direct speaking test followed by topic discussion and information gap tasks rating according to their rating scores. Therefore, the orders of performance and difficulty perception scores on the three test tasks do not completely agree, as the order for the topic discussion and information gap tasks do not concur between the two groups of mean scores.

Table 4 ***Descriptive Statistics of the Items Used in the Correlational Analyses (N = 66)***

	Items	M	SD	Min	Max
Test Score	*Performance of*				
	Pronunciation	5.91	.92	2.80	8.14
	Fluency	5.95	.98	3.26	8.22
	Grammar	5.67	.73	4.00	8.09
	Vocabulary	5.72	.77	3.80	7.37
	Interaction	7.30	1.07	4.31	8.79
	Total	5.98	.84	4.16	8.16
	Performance on				
	Topic discussion	6.04	.87	4.40	8.40
	Information gap	6.15	.93	3.60	8.46
	Semi-direct speaking	5.88	.98	3.53	8.78
Survey response	*Perceived ability of*				
	Pronunciation	3.27	.83	1.00	5.00
	Fluency	3.51	.93	1.00	5.00
	Grammar	3.74	.83	2.00	5.00
	Vocabulary	3.71	.89	2.00	5.00
	Interaction	3.17	.92	1.00	5.00
	Difficulty of				
	Topic discussion	3.97	1.35	1.00	7.00
	Information gap	3.14	1.20	1.00	7.00
	Semi-direct speaking	4.92	1.33	2.00	7.00

Table 5 shows correlations between scores of examinees' perceptions and their performance measured in terms of five English speaking skill components. Only one correlation coefficient was estimated statically significant ($r = .26$, $p < 0.05$) between the two task measures. Other coefficients were non-significant, and their sizes were marginal even including the one which resulted in significance.

Table 5 ***Correlations between Perception and Performance in terms of English Skills***

		Examinee perception				
		1	2	3	4	5
Actual performance	1. Pronunciation	0.21				
	2. Fluency		0.23			
	3. Grammar			0.23		
	4. Vocabulary				0.18	
	5. Interaction					0.26*

* $p < 0.05$

Another correlation matrix is reported in Table 6. The matrix was produced to examine if there was any meaningful relation between examinees' perception of the difficulty of the test tasks and their performance on them. Only one correlation for Information gap was estimated statistically significant ($r = .29$, $p < 0.05$). Especially, the one for the semi-direct speaking test was approaching 0, showing that overall there was no relationship between what examinee did and what they thought about the test. The 0 correlation also informs that there were considerable individual differences about how the examinees regarded the test and how they actually did on the test.

Table 6 ***Correlations between Perception and Performance on the Three Main Tasks***

		Examinee perception		
		1	2	3
Actual performance	1. Topic discussion	0.24		
	2. Information gap		0.29*	
	3. Semi-direct speaking			0.08

* $p < 0.05$

The survey questions for the difficulty of the tasks were not devised to examine the order of the relative task difficulty. Each question was expressed in a 6-point Likert scale, asking how difficult they felt of the concerned task. Therefore, a correlation was estimated to examine the relationship between the orders of performance and difficulty perception on the three tasks. Between the two orders, a very low correlation (r = .06, p > 0.05) was estimated which is close to 0. Such a low correlation informs that there was no meaningful relationship between the order of examinees' performance and that of their perception on the *relative* difficulty of the tasks.

Examinees were also asked to respond to questions in the survey that concerned other aspects of the test tasks. The questions include:

1) Which task do you think measures your English speaking ability best?
2) Which task do you think you did well on?
3) Which task was the most interesting?
4) If you were going to take an English speaking test, which of the three tasks would you prefer most?
5) Which task seems to be most relevant to what you learn in classes at your college?

The response summary to the questions is presented in Table 7. The examinees in this study considered the information gap task as the most desirable method of testing their English speaking ability. They also regarded it as the most interesting and preferred task. Among the three tasks, the examinees thought that they had performed best on the information gap task, but worst on the semi-direct speaking task. Such difficulty perception of the semi-direct task seems to be related to their rating scores, as it received the lowest mean performance rating among the tasks. To the question about the most

relevant tasks to their classroom activities, the examinees selected the topic discussion and information gap tasks as equally close to their classroom tasks. Since the curriculum of the English Language Institute at their college is heavily focused on interactional activities (e.g., group discussions/activities), their answer to this question naturally reflects what they do in their classes.

Table 7 ***Summary of Examinee Responses on the Test Tasks***

Questions	Topic discussion	Information gap	Semi-direct speaking
1. Best English measure	17 (25.8%)	32 (48.4%)	17 (25.8%)
2. Best performed task	17 (25.7%)	39 (59.1%)	10 (15.2%)
3. Most interesting task	9 (13.6%)	42 (63.6%)	15 (22.7%)
4. Most preferred task	9 (13.6%)	40 (60.6%)	17 (25.8%)
5. Most relevant classroom task	27 (40.9%)	29 (43.9%)	10 (15.2%)

Note. Numbers in the question column correspond to the actual questions listed earlier.

Finally, in order to examine if there are any systematic relationships between examinee responses to the questions and their performance, a correlation matrix in Table 8 was examined. Among the correlations, a few were estimated statistically significant. Overall, the examinees seem to think that test tasks must be similar to the tasks they perform in their English classes. Together, the test tasks must be interesting. Considering the information presented in Table 8, they consider the information gap (or topic discussion) task in that regard, but not the semi-direct speaking task.

Table 8 ***Correlations between Perception Variables***

	1	2	3	4
1. Best English measure	--	--	--	--
2. Best performed task	0.08	--	--	--
3. Most interesting task	0.02	0.42**	--	--
4. Most preferred task	0.07	0.35**	0.72**	--
5. Most relevant classroom task	0.31**	0.24*	0.20*	0.15

Note. * $p < 0.05$, ** $p < 0.01$.

IV. CONCLUSION

The current study reported the results from the analyses of survey and rating score data on several oral tasks differing at the structural level. More specifically, it attempted to address the research question of *to what extent the examinees' perceptions of the difficulty are comparable with their performance on three different speech tasks*. A series of questions were devised and given to the examinees in survey format to tap different aspects of examinee perceptions on the tasks. The results of the analyses revealed several interesting facets of the examinee perceptions and their performance on the test-tasks employed in the study.

First, there was no systematic relationship between the examinees' perception on their own proficiency level of the four speaking skill components – pronunciation, fluency, grammar, and vocabulary – and their actual performance of them. The relationship between the Interactional skill and the performance score was an exception. While the interaction is a performance component that requires a more conscious effort, other speaking skills may involve more spontaneous operations, which in turn makes it difficult for the examinees to directly observe and self-evaluate. Such a speculation, however, leads to a pedagogically unfortunate consequence as it may serve as the reason why many of L2 learners find it difficult to improve their English speaking skill(s). Or

as may be the case in this study, different tasks demand different cognitive operations of L2 learners' linguistic resources, while making it difficult for them to allocate their attentional resources to a particular aspect of their speech. In such a case, inherent task features are responsible for the low correlations between examinees' perception on their own proficiency level of the speaking skill components.

Second, there was no meaningful relationship between the order of examinees' perception of the relative difficulty of the tasks and that of their actual performance on them. It may be the very case that perceiving something difficult doesn't necessary mean performing worse on it. Rather, because some task was perceived difficult, some examinees may have paid more attention to the task and tried harder to achieve it. Whichever direction the examinees in this study were taking, it is apparent that there exists much individual directional variability regarding how to approach and deal with task demands. This sort of variability needs to be recognized as a source of invalidity in performance assessment.

Third, the examinees in this study considered the information gap task followed shortly by the topic discussion task as the most suitable testing techniques for their English speaking skills. Many of the examinees felt that they had performed best on the information gap task and worst on the semi-direct speaking task. Their perception on the difficulty of the semi-direct task was related to their rating scores. In addition, the examinees chose both the topic discussion and information gap tasks as similar to what they actually did in their classrooms. This close relation between assessment and instructional tasks renders a supportive argument for the face validity in the use of the interactional tasks rather than the semi-direct tasks.

Learner belief as to the format/technique of assessment has been considered important in language assessment because more psychological attachment to a test task can help ease test anxiety. If that is the case, use of interactional tasks should be more encouraged

even if they may suffer from test reliability. Use of objective criteria and tight training for their use will certainly help enhance the reliability aspect of the performance assessment simultaneously helping to improve face validity.

In conclusion, the finding of this study provides with a converging argument that there is a gap between the examinees' perceived difficulty of the tasks and the level of their performance on them. If the gap is not to be bridged in performance assessment, other types of assessment (e.g., self-assessment using can-do lists) that depend much on learner perception need reconsideration.

REFERENCES

Brindely, G. (2000). Task difficulty and task generalisability in competency-based writing assessment. In G. Brindely (Ed.), *Studies in immigrant English language assessment* (pp. 125-157). Volume 1. Sydney: National Centre for English Language Teaching and Research. Macquarie University.

Brindely, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369-394.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1-25.

Bygate, M. (1999). Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research*, *3*(3), 185–214.

Candlin, C. (1987). Towards task-based language learning. In C. Candlin and D. Murphy (Eds.), *Language learning tasks* (pp. 52). Englewood Cliffs, NJ: Prentice-Hall International.

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.

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

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

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

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett and W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 186-209). Harlow, English: Pearson Educational Limited.

APPENDIX

FOLLOW-UP SURVEY

Name		ID		Class	

1. How would you rate the following English speaking abilities of yours?	
Pronunciation	Very good ------ Good ------ Average ------ Poor ------ Very poor
Grammar	Very good ------ Good ------ Average ------ Poor ------ Very poor
Fluency (an ability to speak smoothly and naturally in English)	Very good ------ Good ------ Average ------ Poor ------ Very poor
Vocabulary	Very good ------ Good ------ Average ------ Poor ------ Very poor
Interaction (an ability to interact with others in English)	Very good ------ Good ------ Average ------ Poor ------ Very poor

2. Have you ever been in an English speaking country?
How long:
When:
For what purpose:

Please answer the following questions.
Please use the following information about the test when you answer the questions below.

Task 1: Discussion about the traditional Japanese family
Task 2: Discussion to decide a present for John in Hawaii
Task 3: Speaking test using a computer; took this alone

1. Tasks 1~3 の難易度を「1 非常に簡単」から「6 非常に難しい」まで 6 段階で評価してください。						
	(非常に簡単)		(やや簡単)	(やや難しい)		(非常に難しい)
Task 1	1 ----------	2 ----------	3 ----------	4 ----------	5 ----------	6
Task 2	1 ----------	2 ----------	3 ----------	4 ----------	5 ----------	6
Task 3	1 ----------	2 ----------	3 ----------	4 ----------	5 ----------	6

2. Which task was the easiest to take?
Put 1 next to the task that was the easiest, 2 next to he second easiest , and 3 next to the most difficult.

Task 1 _____ Task 2 _____ Task 3 _____

3. Which task do you think measures your English speaking ability best?
Put 1 next to the task that would measure best, 2 next to the second best, and 3 next to the worst task.

Task 1 _____ Task 2 _____ Task 3 _____

4. Which task do you think you did well?
Put 1 next to the task that you did best, 2 next to the second best, and 3 next to the task you did worst.

Task 1 _____ Task 2 _____ Task 3 _____

5. Which task was the most interesting?
Put 1 next to the most interesting task, 2 next to the second most interesting, and 3 next to the least interesting task.

Task 1 _____ Task 2 _____ Task 3 _____

6. If you were going to take an English speaking test, which of the three tasks would you prefer most?
Put 1 next to the task that you prefer most, and 2 to the second, and 3 to the task you prefer least.

Task 1 _____ Task 2 _____ Task 3 _____

7. Which task seems to be most relevant to what you learn in classes at KUIS?
Put 1 next to the task most relevant to your class activities, 2 to the second most related task, and 3 to the task that is related least to the class activities.

Task 1 _____ Task 2 _____ Task 3 _____