『コーパスに基づく言語学教育研究報告』No.9（2012）

Test-takers' affective reactions to a computer-delivered speaking test and their test performance

Yujia ZHOU
(Tokyo Tech University)

Abstract

Few studies have focused on test-takers' affective reactions to computer-delivered speaking tests, and little is known about the relationship between these reactions and their actual test performance. This study investigated the affective reactions of Japanese learners to a computer-delivered speaking test and the relationship of these reactions with the test-takers' proficiency level and test performance. Seventy-six Japanese learners completed a questionnaire that elicited their affective reactions in four dimensions after taking a computer-delivered speaking test. The results indicate that the participants had mixed reactions: they felt nervous in taking the test and considered the test difficult, but enjoyed the test taking experience and perceived the test as fair and valid. Two factors—liking for the test and perceptions of test difficulty—differentiated the low-proficiency group from the high-proficiency group. These two factors were also the best predictors of the participants' test performance and accounted for about 30% of the variance in test performance. The participants who showed more favorable emotions to the test and who perceived the test as less difficult tended to obtain higher scores.

1. Introduction

In recent years, computers have been increasingly used in testing second language speaking skills. Several large-scale test programs such as the Test of English as a Foreign Language Internet-based test (TOEFL iBT) and Test of English for International Communication (TOEIC) have launched computer-delivered speaking tests. In these tests, the test-taker talks to a computer instead of an interlocutor. This has many advantages including cost reduction and increased efficiency in test administration, the rating process, and the score report.

However, the absence of the interlocutor poses many potential problems for test validity. One of the key issues lies in test-takers' affective reactions to computer-delivered speaking tests, which can threaten several aspects of test validity, namely, face validity, test fairness, and score interpretation. First, since oral communication usually involves interactions between people, computer-delivered

speaking tests seem to lack face validity in terms of authenticity. If this leads to negative affective reactions among test-takers, they might not take the test seriously or respond appropriately to test tasks. Second, test-takers at different proficiency levels may react differently to testing speaking using computers. If this were to occur, the fairness of the test would be compromised. Finally and most importantly, if the test-takers' affective reactions were systematically related to test performance, the scores on a computer-delivered speaking test might reflect not only their speaking ability, but also the level of their affective reactions. As such, construct-irrelevant variance would be introduced into the measure, which would confound the interpretation of the test scores (Messick, 1989).

Despite the importance of these issues, few studies have collected baseline information about test-takers' affective reactions to the computer-delivered speaking test, and little is known about the relationship between these reactions and their actual test performance. This study reports a study of Japanese English-as-a-foreign-language (EFL) learners' affective reactions to a computer-delivered speaking test developed in Japan and the relationship of these reactions with the test-takers' proficiency level and test performance. The findings of the study will provide empirical evidence for potential concerns about face validity, test fairness, and score interpretations of the computer-delivered speaking test.

1.1. Affective reactions to computer-delivered speaking tests

Although there is evidence from a few studies that test-takers generally show positive attitudes towards computer-delivered speaking tests, their reactions were not so favorable when compared to the reactions to interview tests or computer-based language tests of other skills. In Sapsirin (2007), university students in Thailand considered the computer-delivered speaking test used in the study to be a good measure of their speaking ability and the computer to be a useful and appropriate tool for delivering a speaking test. They also indicated that taking the test was a pleasant experience because they felt less stressful and more confident. However, comparing a computer-delivered speaking test with a face-to-face version of the test, Zhou (2009) revealed that Japanese learners preferred the face-to-face test, which was perceived to be more pleasant and more accurate in assessing their spoken English. In Stricker and Attali (2010), the responses of test-takers to one questionnaire item on the TOEFL iBT—*The TOEFL gave me a good opportunity to demonstrate my ability to speak English*—were consistently less favorable in four countries than their responses in other sections.

Given the fact that computer-based speaking tests are being widely used, more information about the affective reactions of test-takers needs to be accumulated. This type of information would provide evidence for the face validity of computer-delivered speaking tests and would also be useful to test developers in designing test tasks.

1.2. Affective reactions to computer-delivered speaking tests and proficiency level

It is possible that the affective reactions to a computer-delivered speaking test would not be

uniform across respondents. This study is interested in whether the test-takers at different proficiency levels would react differently to a computer-delivered speaking test.

It can be hypothesized that higher-proficiency learners have more positive reactions to a computer-delivered speaking test than lower-proficiency learners do. The rationale is that in a computer-delivered speaking test, lower-level learners might be at a disadvantage for two reasons. First, test-takers with insufficient proficiency are likely to feel less assured when talking to a computer without an interlocutor present to nod to them and to give at least minimal responses. However, higher-proficiency learners may not be influenced as much by the absence of an interlocutor. Second, a computer-delivered speaking test may be better suited to elicit extended speech and consequently, may be friendlier to higher-proficiency learners. On the other land, a face-to-face test may be more suitable for lower-proficiency learners, as it involves short responses that are more typical of interactions between two people.

Most previous studies have provided empirical evidence in support of the view that proficiency level is related to the affective reactions of test-takers. Bradshaw (1990) found that the high-proficiency group perceived the C-test to be more valid and more interesting and the reading test to be a fairer test than the low-proficiency group did. Similarly, Scott and Madsen (1983) show that learners with low levels of proficiency rated interview tasks less favorably than more proficient learners did. In Kenyon and Malabonga (2001), test-takers of the low proficiency level reported the Simulated Oral Proficiency Interview (SOPI), a tape-based speaking test, to be more difficult than the other two higher proficiency groups did. On the contrary, Sapsirin (2007) observed no significant differences in the affective reactions to the computer-delivered speaking test among three proficiency groups. However, the lack of differences could be attributed to the fact that the participants in the study were all university students, who could be considered to represent only a narrow range of proficiencies.

1.3. Affective reactions to computer-delivered speaking tests and test performance

Krashen's (1985) Affective Filter Hypothesis provides theoretical support for the possible relationships between the affective reactions of test-takers and their test performance. Krashen proposes that the existence or removal of an affective filter might directly affect a learner's learning outcome of a second or foreign language. Consequently, an affective filter might also be at work in a testing situation, and the emotional reactions of test-takers could affect their test result. This is a particularly important issue in the context of the computer-delivered speaking test, since the absence of an interlocutor seems to be unnatural for daily communication.

Previous studies found that favorable affective reactions to language tests were often significantly related to the higher performance. However, the existing research does not show a clear relationship between each dimension of affective reactions and test performance; this could be due to the different tests used and the different questionnaire items employed. Test-taking anxiety tends to be negatively associated with learners' performance in face-to-face interviews (Young, 1986; Phillips,

1992) and in group and pair tests of speaking (Scott, 1986). The higher the test-takers' test score, the less difficult they found the tape-based speaking test (Brown, 1993) or specific test tasks (Elder, Iwashita, & McNamara, 2002). On the other hand, neither the enjoyment in taking a tape-based speaking test (Brown, 1993; Elder et al., 2002) nor the perceptions of test validity (Brown, 1993) were found to significantly correlate with the test scores.

As for the speaking test delivered using a computer, to the best of the author' knowledge, no empirical evidence is available on how test-takers' reactions are related to test performance. Stricker and Attali (2010) examined the relationships between the test takers' general acceptance of the TOEFL iBT and the test scores in each of the four sections. The study mainly elicited the test takers' perceptions of test difficulty and the test validity of the TOEFL iBT test as a whole. The general acceptance of the test was found to significantly correlate to the test scores on all the sections except the speaking section. Since the speaking section is only one part of the test, it is difficult to interpret the results. Nevertheless, the results seem to suggest the need to investigate the relationship between speaking test scores and test takers' reactions specific to the speaking section.

In sum, while researchers seem to agree that learners with positive reactions to a test tend to have higher scores, there are no conclusive results regarding the dimensions that are related to test performance. Investigations on the various dimensions of affective reactions could shed more light on this issue. Methodologically, all the studies discussed so far used single items to represent each affective dimension. Given that psychological factors are usually multi-dimensional, using scale-level items may be more appropriate in measuring affective factors.

2. The present study

Studies on the affective reactions of test-takers to computer-delivered speaking tests are still few in number, and none has linked affective reactions to the exploration of test performance. Thus, the purpose of the present study was to explore the affective reactions of test-takers to a computer-delivered speaking test and the relationships of these reactions with the test-takers' proficiency level and test performance. The dimensions of affective reactions investigated were test-taking anxiety, liking for the test, perceptions of test difficulty, and perceptions of test validity. Further, in order to reflect the multi-dimensional nature of each construct, several items were used to represent each dimension. Specifically, this study investigated three research questions:

(1) What are the affective reactions of Japanese EFL learners to the computer-delivered speaking test?
(2) What affective reactions to the computer-delivered speaking test differentiate Japanese EFL high-proficiency learners from low-proficiency learners?
(3) To what extent are the affective reactions of Japanese EFL learners to the computer-delivered speaking test related to their test performance?

3. Method

3.1. Participants

Seventy-six Japanese EFL students participated in this study. The students were from three universities (81%) and two high schools (19%) in the district of Tokyo. There were 23 male (24%) and 73 female (76%) students. Of the undergraduate students, 34 students from a foreign language university were specializing in foreign languages other than English. Ten female students from a comprehensive university were majoring in English language and literature. The 34 female undergraduate students majoring in nursing and housekeeping were from a women's university. Of the 18 high school students, ten male students were from a boys' high school, and one male student and seven female students were from a coeducational high school. All the students participated on a voluntary basis and received an honorarium for their participation.

3.2. Computer-delivered speaking test

For this study, the speaking part of the Global Test of English Communication (GTEC) for STUDENTS was used[1]. The test, developed by the Benesse Corporation in Japan, is a four-skill, computer-based English test. It mainly targets Japanese high school and university students. The speaking part of the GTEC for STUDENTS consists of four sections. The first section includes a read-aloud task that involves the reading of a 40–word paragraph out loud. The second section involves a repetition task with three items that vary in grammatical complexity. Section Three contains four pictures that tell a simple story; the participants are given two minutes to narrate the story. The last section—the opinion task—provides a graph, and the participants are required to give their opinions about a topic based on the information in the graph. The whole test lasts about 15 minutes.

Both the oral and written instructions are in Japanese. In the first three sections, video prompts are presented, in which an American lady performs the role of asking questions and giving simple, preset feedback in English. The test-takers were given some time to prepare their responses in each section. They could begin recording their responses by clicking the start button on the screen when they were ready, and their responses were recorded automatically when the preparation time was over. Note taking was not allowed during the test.

Each task was scored by sets of two accredited raters on a differential combination of rating scales for pronunciation, fluency, grammar, and vocabulary on a 0–4 scale. The inter-rater reliability estimates for the ratings on each rating element were moderate, ranging from .52 to .75. The scores for each element of each task were determined by taking the mean of the ratings from the two raters. The final scores for each rating element were calculated by averaging the element scores of each task; these were added up to obtain a total score for each participant.

[1] The speaking part of the GTEC for STUDENTS is not in use as of 2009.

3.3. Questionnaire

A questionnaire in English was designed for this study to measure the test-takers' affective reactions to the speaking part of the GTEC for STUDENTS. The questionnaire was adapted from several studies including Scott (1986), Brown (1993), Keynon and Malabonga (2001), and Sapsirin (2007). It was translated into Japanese and was verified by a native speaker of Japanese who specialized in applied linguistics. It consisted of 15 items that elicited the test-takers' affective reactions to the computer-delivered speaking test on four dimensions: test-taking anxiety, liking for the test, perceptions of test difficulty, and perceptions of test validity (see Table 1). The test-takers were asked to indicate the degree to which they agreed with each item using a five-point Likert scale, where 5 = *Strongly agree*, 4 = *Agree*, 3 = *Neutral*, 2 = *Disagree*, and 1 = *Strongly disagree*.

3.4. Data collection procedures

Data were collected between November and December 2006. Each university student took the speaking part of the GTEC for STUDENTS alone in a quiet seminar room or with about 20 other classmates in their computer laboratory. All the high school students took the test in a computer laboratory together with two or three other classmates. They were made to sit far apart so that they would not influence one another. Immediately after the test, the participants were asked to complete and submit the questionnaire to their instructors. The ratings for each task on each rating element were obtained from the Benesse Corporation directly.

3.5. Data analysis

First, a preliminary data analysis was conducted to screen the data by examining the descriptive statistics. The Pearson product-moment correlation matrix was computed to assess the multicollinearity of the data. Following Field (2005), a correlation of above .90 between two variables was considered a threat to the singularity of the data. Further, an exploratory factor analysis was conducted to verify whether the results were in accordance with the hypothesized structures that were assumed to underlie the questionnaire. The principal component factoring with direct oblimin rotation (varimax) was used in the factor analysis.

To address the first research question, that is, —to investigate the Japanese EFL learners' affective reactions to the speaking part of the GTEC for STUDENTS, the descriptive statistics were analysed at the scale level. To address the second research question, that is, —to explore the relationships between the affective factors and the test-takers' proficiency level, all the participants were divided into two groups according to the median of their test scores. Independent-sample *t*-tests were conducted, with the mean scores of the factor scales as the dependent variable. A multiple regression analysis was performed to explore the third research question, that is, —to examine the relationships between the affective factors of test-takers and their test performance. SPSS 12.0 was used to conduct all the data analysis. An alpha level of .05 was used for all tests of significance.

4. Results

4.1. Data screening

Of the 76 questionnaires that were returned, only one had missing data. This case was removed from the dataset, leaving 75 valid responses. Table 1 presents the descriptive statistics of all the 15 items, which reveal that all the data, except for item 8, were normally distributed with values for skewness and kurtosis within an acceptable range of –2 to +2. Since item 8 had a skewness of and kurtosis of –1.72 and 3.24, respectively, this variable was not considered for further analysis. The Pearson product-moment correlation figures among all the variables were found to be no larger than .74, indicating that multicollinearity was not a problem for the present data.

Table 1: Descriptive statistics of all the questionnaire items

No	Items	*M*	*SD*	*Skewness*	*Kurtosis*
Test-taking anxiety					
1	I felt nervous before the test.	3.36	1.30	-0.33	-1.09
2	I felt nervous when I was taking the test.	3.25	1.27	-0.29	-1.18
3	I would have performed better if I had not got nervous.	2.97	1.09	-0.01	-0.85
Liking for the test					
4	I like the format of the test.	3.04	1.01	-0.16	-0.15
5	The test was interesting.	3.51	1.17	-0.72	-0.31
6	Taking the speaking test on computer was not a pleasant experience.	2.47	1.06	0.23	-0.60
7	I am used to the format of the test.	1.81	0.82	0.82	0.22
Perceptions of test difficulty					
8	I believe I did well on the tasks.	1.68	0.93	1.72	3.24
9	I felt confident when I did the test.	1.80	1.01	1.22	0.73
10	I felt the test was difficult.	3.59	1.07	-0.40	-0.29
11	I would have performed better had the format of the test been different.	2.64	0.88	0.17	0.33
12	I would have performed better if I had known the topics of the test tasks better.	3.09	0.98	0.08	-0.70
Perceptions of test validity					
13	I believe the format and the content of the test was fair.	3.59	0.95	-0.36	-0.32
14	I believe I had enough opportunity to show my ability to speak English.	2.60	1.14	0.39	-0.54
15	The test reflects accurately how well I speak English.	3.16	0.93	0.09	-0.21

4.2. Exploratory principal component analysis

The exploratory factor analysis confirmed the distinctiveness of the factor structure of the four dimensions. The scree plot indicated that a four-factor model was the most appropriate for the data

and accounted for about 62.12 % of the total variance.

Table 2 shows the factor loadings of the items for each factor. A factor loading above .50 for a particular factor was the criterion used for selecting the items for each factor. The rationale for this criterion is that although typically, researchers take a loading of an absolute value more than .30 to be important, the significance of a factor loading depends on the sample size. According to Stevens (1992), for a sample size of 50, a loading of .72 can be considered significant, and for a sample size of 100, the loading should be greater than .51. Since the sample size of this study was 75, it was decided to include only those items with a factor loading above .50. Accordingly, items 14 and 7 were deleted.

Table 2: Factor loadings of questionnaire items

No.	Items	Factor 1	Factor 2	Factor 3	Factor 4
1	I felt nervous before the test.	**0.90**	0.01	0.00	-0.06
2	I felt nervous when I was taking the test.	**0.81**	-0.03	0.31	-0.12
3	I would have performed better if I had not got nervous.	**0.78**	0.02	-0.04	-0.13
14	I believe I had enough opportunity to show my ability to speak English.	-0.49	0.39	-0.08	-0.20
4	I like the format of the test.	0.09	**0.79**	0.05	0.16
5	The test was interesting.	0.14	**0.78**	-0.16	0.35
6	Taking the speaking test on computer was not a pleasant experience.	0.17	**-0.73**	0.00	-0.25
7	I am used to the format of the test.	-0.16	0.43	-0.34	-0.07
10	I felt the test was difficult.	0.02	-0.16	**0.76**	0.09
12	I would have performed better if I had known the topics of the test tasks better.	0.05	0.08	**0.74**	-0.08
9	I felt confident when I did the test.	-0.23	0.45	**-0.54**	-0.32
15	The test reflects accurately how well I speak English.	-0.03	0.26	0.05	**0.75**
11	I would have performed better had the format of the test been different.	0.40	-0.13	-0.22	**-0.67**
13	I believe the format and the content of the test was fair.	-0.01	0.13	-0.47	**0.61**
	Variance explained (%)	18.68	17.43	13.14	12.88

As shown in Table 2, the first three factors all contained three items that were intended to measure the original dimensions. Therefore, these factors retained their original names: “Test-anxiety” (Factor 1), “Liking for the test” (Factor 2), and “Perceptions of test difficulty” (Factor 3). However, item 11 from the “Perceptions of test difficulty” scale —*I would have performed better had the format of the test been different*—had a moderately high loading for Factor 4. Since this item could also be interpreted as eliciting responses regarding the fairness of the test,

Factor 4 also retained its original name, “Perceptions of test validity”. The variance explained by the four factors was 18.68%, 17.43%, 13.14%, and 12.88%, respectively.

The internal-consistency reliabilities were satisfactory for the “Test-taking anxiety” scale (α = .83) and for the “Liking for the test” scale (α = .79), but were moderate for the “Perceptions of test difficulty” scale (α = .58) and for the “Perceptions of test validity” scale (α = .59). The scale scores were obtained for each affective factor by adding up the responses for all the items of that particular scale.

4.3. Affective reactions to the computer-delivered speaking test

Descriptive statistics were conducted to demonstrate the affective reactions of the test-takers to the computer-delivered speaking test at the scale level. As shown in Table 3, when considering the participants as a single group, the mean scores of the scales were all above three, indicating that the test-takers generally agreed with the statements included in each scale. Factor 3, “Perceptions of test difficulty”, received the strongest agreement (M = 3.63, SD = 0.75) across the test-takers, followed by Factor 4, “Perceptions of test validity” (M = 3.37, SD = 0.68), Factor 2, “Liking for the test” (M = 3.36, SD = 0.90), and Factor 1, “Test-taking anxiety” (M = 3.20, SD = 1.06). These results suggest that the test-takers agreed that the test was rather difficult, and they perceived the test to be fair and valid for testing their speaking ability. They also reported that they felt nervous before and during the test. Finally, they liked the format of the test and enjoyed taking the test.

Table 3: Descriptive statistics and independent *t*-test of low-proficiency and high-proficiency learners’ affective reactions

Factor	*Prof*	*M*	*SD*	*N*	*p*
1 Test-taking anxiety	Low	3.13	1.16	37	0.58
	High	3.26	0.96	38	
	Total	3.20	1.06	75	
2 Liking for the test	Low	3.05	1.01	37	0.00
	High	3.66	0.67	38	
	Total	3.36	0.90	75	
3 Perceptions of test difficulty	Low	3.95	0.69	37	0.00
	High	3.31	0.68	38	
	Total	3.63	0.75	75	
4 Perceptions of test validity	Low	3.38	0.75	37	0.91
	High	3.36	0.61	38	
	Total	3.37	0.68	75	

Note: *Prof* = proficiency groups.

4.4. Affective reactions and proficiency level

In order to evaluate the differences in the test-takers' affective reactions due to their proficiency levels, an independent *t*-test was carried out. The mean scores of each scale were compared for the 38 high-performing test-takers who scored above the median, and 37 low-performing test-takers who scored below the median.

As indicated in Table 3, the responses of the low-proficiency group on two of the affective factors, namely, "Test-taking anxiety" and "Liking for the test," were slightly lower than those of the high-proficiency group were. For Factor 3, "Perceptions of test difficulty" and Factor 4, "Perceptions of test validity", the low-proficiency group's ratings were slightly higher. Further, there were significant differences between the two proficiency groups in two scales: "Liking for the test" and "Perceptions of test difficulty". The low-proficiency group reported the test to be more difficult (M = 3.95, SD = 0.69) than the high-proficiency group did (M = 3.31, SD = 0.68). The high-proficiency group showed more favorable emotions towards the test (M = 3.66, SD = 0.67) than the low-proficiency group did (M = 3.05, SD = 1.01).

However, "Test-taking anxiety" and "Perceptions of test validity" were not found to be significantly different between the two proficiency groups. That is, the test-takers in the two proficiency groups did not feel different degrees of nervousness, and both the groups considered the test fair and valid.

4.5. Affective reactions and test performance

A stepwise multiple regression analysis was performed to determine the relationships between the affective reactions of the test-takers to the computer-delivered speaking test and their test performance. The dependent variable was the total score of the test, and the independent variables were the scale scores of the four affective factors.

As shown in Table 4, Model 2— the final model adopted— included two predictors, "Perceptions of test difficulty" and "Liking for the test". However, "Test-taking anxiety" and "Perceptions of test validity" did not emerge as statistically significant predictors. The R^2 of the model was .30, which indicates that this model was able to explain 30% of the variance in the test-takers' performance on the test.

The standardized coefficients in Table 4 demonstrate that "Perceptions of test difficulty" had a slightly stronger relationship with the test score (β = –.40), followed by "Liking for the test" (β = .30). The results suggest that the test-takers who felt the test was more difficult tended to have lower test scores. The test-takers who held more favorable emotions towards the test tended to obtain higher scores.

Table 4: Multiple regression analysis on test takers' affective reactions to the computer-delivered speaking test and test performance

Model	Predictors	β	t	p	R^2
1	(Constant)		11.17	0.00	0.22
	Perceptions of test difficulty	-0.47	-4.52	0.00	
2	(Constant)		6.26	0.00	0.30
	Perceptions of test difficulty	-0.40	-4.00	0.00	
	Liking for the test	0.30	2.96	0.00	

5. Discussion

5.1. Research Question 1: What are the affective reactions of Japanese EFL learners to the computer-delivered speaking test?

The descriptive statistics at the scale level revealed the mixed reactions of the test-takers to the computer-delivered speaking test. The test-takers had negative reactions to the test in respect of test-taking anxiety and perceptions of test difficulty but had positive reactions in terms of perceptions of test validity and liking for the test.

In this study, the test was administered in an experimental setting and had no impact on the test-takers' academic records; yet, the participants reported a certain degree of nervousness in taking the test. One possible explanation for this result is that their nervousness might have been caused by their unfamiliarity with the format of the test. This is supported by their low ratings for the questionnaire item *–I am used to the format of the test* (M = 1.81). In fact, high school students as well as university students in Japan receive limited exposure to speaking tests in general, let alone computer-delivered speaking tests. Moreover, they do not use computers frequently in their daily life. The combination of these factors may have led to uncertainty in the minds of the test-takers, which in turn created stress.

This study also found that the test-takers perceived the test to be rather difficult. This result could also be ascribed to the unfamiliarity of the test-takers with the computer-delivered speaking test. Further, the test-takers may have felt that a particular task type was more difficult than the others owing to various factors such as time limits, difficult vocabulary, speed of the native speakers' speech in the recordings, and lack of clarity in the instructions. As reported in McNamara (1990), some of the negative reactions to the tape-based speaking test used in the study appeared not to be a reaction to the tape-based test per se, but a reaction to the type of the task set. However, we have no way to determine this conclusively, since the test-takers were not asked to give responses or to write specific comments concerning each task. Future research would need to explore task-specific responses in order to gain more insights about this issue.

Another finding of the study was that despite their negative reactions, the participants perceived the test as fair and believed that it reflected their speaking ability. This indicates that the test-takers were able to make objective judgments regarding the test, and it also provides positive evidence for

the face validity of the computer-delivered speaking test. It was also interesting to find that although the test-takers felt fairly nervous and perceived the test as rather difficult, they still reported that they enjoyed taking the test. A plausible explanation is that the test-takers' positive perceptions of test validity influenced their decisions about how well they liked the test. As discussed in Savignon (1972), the students reacted very positively to the test even though it was difficult for them because they felt it actually tested the language skills they were trying to acquire. Another viable interpretation lies in the profile of the test-takers in this study. They volunteered for the study, and their instructors noted that they had higher motivation to learn English. They may have felt satisfied with getting the opportunity to speak English during the test, and therefore, they enjoyed the experience of taking the test.

5.2. Research Question 2: What affective reactions to the computer-delivered speaking test differentiate Japanese EFL high-proficiency learners from low-proficiency learners?

The *t*-test results show that liking for the test and perceptions of test difficulty were the two important factors that differentiated the high-proficiency and low-proficiency test-takers. The higher-proficiency group reported the test to be less difficult, and they thought taking the test was a more pleasant experience than the lower-proficiency test-takers did. These findings support our hypothesis that the higher level test-takers would have more positive affective reactions to the computer-delivered speaking test than the lower level test-takers, and they are in line with the findings of Bradshaw (1990), Scott and Madsen (1983), and Kenyon and Malabonga (2001).

However, this study did not reveal significant differences between the two proficiency groups with respect to test-taking anxiety and perceptions of test validity. These findings contradicted our hypothesis that the higher-proficiency group would hold more positive reactions to the computer-delivered speaking test than the lower-proficiency group. There are two possible explanations for the lack of differences between the two groups. First, this may be due to the limited range of the proficiency levels of the participants in this study. High school students were involved in order to represent a wider range of proficiencies of Japanese EFL learners. However, the high school volunteers had higher motivation and performed better than their classmates did, as was pointed out by their instructors. Thus, it is possible that the oral proficiency of the test-takers assigned to the lower level was not low enough. Future studies may need to recruit participants from among high school students in lower grades or junior high school students. Second, it is important to note that the criterion for grouping the participants was the scores obtained from the ratings of the tasks. Using a speaking proficiency criterion from another established speaking test might produce a different outcome.

5.3. Research Question 3: To what extent are the affective reactions of Japanese EFL learners to the computer-delivered speaking test related to their test performance?

The multiple regression analysis revealed that two dimensions of the affective reactions of test-takers were linked with their performance on the computer-delivered speaking test: liking for the test and perceptions of test difficulty. The test-takers who showed more favorable emotions toward and more interest in the computer-delivered speaking test tended to have higher scores. Those who considered the test more difficult tended to obtain lower scores. These findings are consistent with Krashen (1985)'s Affective Filter Hypothesis and the findings reported in previous studies (Brown, 1993; Elder et al., 2002). These two aspects of the test-takers' affective reactions explained 30% of the variance in the test scores. It is important to note that in addition to the speaking ability that was the target of the test, the variance in the speaking test score could be influenced by a host of factors including the test method, task factors, rating scale, or other test-taker characteristics. In this regard, the findings of the present study suggest that the affective reactions of test-takers can be considered an important source of construct-irrelevant variance that requires our attention in score interpretation.

In contrast, test-taking anxiety seems not to be related to test performance. This is not in line with previous research which suggests that high test-taking anxiety leads to underachievement. One possible explanation for the discrepancy in the findings is that different types of speaking tests were used in these studies. Previous studies were based on face-to face tests that were either group oral (Scott, 1986) or interview-based (Young, 1986; Philips, 1992), whereas this study involved a computer-delivered test with no interlocutor. Thus, a tentative suggestion would be that test-taking anxiety does not predict test performance on computer-delivered speaking tests. However, more research evidence is needed to corroborate this proposition. Another possibility is that test-taking anxiety did have an effect on test performance, but it was an indirect one through the mediation of factors such as time constraint. Hill (1983) observed that under time pressure, high test-anxious students made three times as many errors as low test-anxious students. However, when the time limits were removed, the high test-anxious students performed as well as their low test-anxious peers did. In the present case, the test-takers were given one minute and two minutes to respond to the story-telling task and the opinion task, respectively. Probably because of their limited oral proficiency, most of the test-takers were observed to give only short responses, and they remained silent for quite a long period, particularly during the opinion task. That is, the response time provided may not have functioned as a time constraint for the participants. Thus, had the response time on the monologic tasks been shorter or had the tasks been easier, a significant relationship between test-taking anxiety and test performance might have been detected.

The findings of this study also suggest that the perceptions of test validity seem not to predict performance on the computer-delivered speaking test. It can be argued that perceptions of test validity may have an indirect effect on test performance through the mediation of the other two factors: liking for the test and motivation. As discussed earlier, test validity was an important factor

in the students' decisions about how well they liked the test. Given that liking for the test was found to be a significant factor in affecting test performance, perceptions of test validity may have an indirect effect on test performance through the mediation of liking for the test. It is also possible that perceptions of test validity affected performance through motivation. The test-takers who considered the test more fair and valid may have had higher motivation to achieve better performance. Future research could shed more light on this issue by including the variable of motivation in the questionnaire.

6. Conclusion

This study explored how Japanese EFL learners reacted to a computer-delivered speaking test, and how these reactions were related to the test-takers' proficiency level and their performance on the speaking test. The findings show that the test-takers had mixed reactions to the test. While they felt a certain degree of test-taking anxiety and considered the test rather difficult, they did enjoy the experience of taking the test and perceived the test to be fair and valid. Their positive reactions, especially those regarding the test validity, provide positive evidence for the face validity of the test. It seems that the test is acceptable to the prospective test-takers. It was also found that liking for the test and perceptions of test difficulty were the two important factors that differentiated the high-proficiency and low-proficiency test-takers. Further, these affective reactions of the test-takers were significantly related to their performance on the computer-delivered speaking test. These findings suggest that the test is biased towards test-takers at different proficiency levels in terms of affective reactions. In addition, these reactions represent an important source of irrelevant variance in test performance, which provides empirical evidence for the effect of affective factors on the test-takers' performance. Accordingly, we need to be careful about the interpretation of the scores on computer-delivered speaking tests. Overall, the results of this study would be of interest not only to researchers, but also to the test developers who would need to ensure that the computer-delivered speaking test is acceptable to the prospective test-takers and fair to all the candidates.

We need to be cautious when generalizing the results to test-takers of other gender groups and other nationalities, as well as to other computer-delivered speaking tests. First, prior research has found significant differences in the affective reactions of male and female participants to speaking tests. Bensoussan and Zeidner (1989) show that women reacted more negatively towards oral language tests and experienced more anxiety than men. Hill (1998) reported that women felt the live interview to be the more difficult format more often than was expected. Considering that the majority of this study's participants were female, the results would have been varied if the percentage of men and women had been different. Second, learners with different cultural backgrounds seem to have different affective reactions to language tests. Scott (1980) found that Spanish students evaluated the English proficiency test batteries more positively than Japanese students did. Similarly, Stricker and Attali (2010) reported that the test-takers' attitudes towards the speaking section of the TOEFL iBT varied markedly by country. Finally, the test-takers' reactions

to the computer-delivered speaking test could be specific to the context of this study. In order to generalize the findings, we would need to take into account the features of the test used in this study as well as contextual factors that may include the test-takers' attitudes towards English-learning environment, the role of English in achieving goals, and test practice in and outside the classroom.

Given that the application of computer technology for testing speaking skills is still in the initial stages, further investigations are required to add to the findings of the present study. Continuous efforts should be made to improve the reliability of the questionnaire used. Using a larger number of items that represent each dimension more accurately would help to improve the reliability of each scale. A vital area that future research should focus on is the interrelationships among the affective factors, in order to improve our understanding of how they interact in affecting the performance of test-takers. Such studies would entail the use of multivariate techniques such as structural equation modeling.

Acknowledgements

I am grateful to the two anonymous reviewers for their valuable suggestions in revising the article. I acknowledge the contribution of the following people in data collection: Negishi Masashi, Asako Yoshitomi, Naoyuki Kiryu, Asako Kaneko, Atushide Tanaka, Hitoshi Sugawara, Yukio Tono, and Takahiro Kowata. Finally, I thank Benesse Corporation for giving permission to use the speaking part of the GTEC for STUDENTS.

References

Bensoussan, M., & Zeidner, M. (1989) Anxiety and achievement in a multicultural situation: The oral testing of advanced reading comprehension. *Assessment and Evaluation in Higher Education*, 14(1) 40–54.

Bradshaw, J. (1990) Test-takers' reactions to a placement test. Language Testing, 7(1) 13–30.

Brown, A. (1993) The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10(3) 277–303.

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

Field, A. (2005) *Discovering Statistics Using SPSS*. London: Sage.

Hill, K. T. (1983) Interfering effects of test anxiety on test performance: A growing educational problem and solutions to it. *Illinois School Research and Development*, 20(1) 8–19.

Hill, K. T. (1998) The effect of test-taker characteristics on reaction to and performance on an oral English proficiency test. In A. J. Kunnan, editor, *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 209–229). Mahwah, NJ: Lawrence Erlbaum Associates.

Kenyon, D. M., & Malabonga, V. (2001) Comparing examinee attitudes toward computer-assisted

and other oral proficiency assessments. *Language Learning & Technology*, 5(2) 60–83.

Krashen, S. (1985) *Input Hypothesis: Issues and Implications*. Longman.

McNamara, T. F. (1990) *Assessing the Second Language Proficiency of Health Professionals*. Unpublished doctoral dissertation, University of Melbourne.

Messick, S. (1989) Validity. In R. L. Linn, editor, *Educational measurement* (3rd ed.) (pp.13–103). New York: American Council on Educational/Macmillan.

Phillips, E. (1992) The effects of language anxiety on students' oral performance and attitudes. *The Modern Language Journal*, 76(1) 14–25.

Sapsirin, S. (2007) *A Study of Trait Factors of Oral Language Ability in a Computer-Based Speaking Test for Thai University Students*. Unpublished doctoral dissertation, Chulalongkorn University, Thailand.

Savignon, S. J. (1972) *Communicative Competence: An Experiment in Foreign Language Teaching*. Philadelphia: Center for Curriculum Development.

Scott, M. L. (1980) *The Effect of Multiple Retesting on Affect and Test Performance*. Unpublished MA thesis, Brigham Young University.

Scott, M. L. (1986) Student affective reactions to oral language tests. *Language Testing*, 3(1) 99–118.

Scott, M. L., & Madsen, H. S. (1983) The influence of retesting on test affect. In J. W. Oller, editor, *Issues in language testing research* (pp. 270–279). Rowley, MA: Newbury House.

Stevens, J. P. (1992) *Applied Multivariate Statistics for the Social Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Stricker, L., & Attali, Y. (2010) *Test Takers' Attitudes about the TOEFL iBT* (TOEFL iBT Research Report No. 13). Princeton, NJ: Educational Testing Service.

Young, D. J. (1986) The relationship between anxiety and foreign language oral proficiency rating. *Foreign Language Annals*, 12(5) 439–488.

Zhou, Y. J. (2009) *Effects of Computer Delivery Mode on Testing Second Language Speaking: The Case of Monologic Tasks*. Unpublished doctoral dissertation, Tokyo University of Foreign Studies, Japan.