Zhou *Language Testing in Asia* (2015) 5:2
DOI 10.1186/s40468-014-0012-y

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Computer-delivered or face-to-face: effects of delivery mode on the testing of second language speaking

Yujia Zhou

Correspondence:
zyujia2400@yahoo.co.jp
The Faculty of Foreign Studies,
Tokyo University of Foreign Studies,
3-11-1, Asahi-cho, Fuchu-shi, Tokyo
183-8534, Japan

Abstract

Background: The use of computers has increased in speaking assessments; however, there are concerns about how the absence of an interlocutor affects performance on speaking tests.

Method: In the current study, the test scores and the underlying factor structures of monologic tasks were compared between two delivery modes: computer delivery and face-to-face modes. Seventy-nine Japanese students responded to two monologic tasks delivered via both modes in a counterbalanced design.

Results: One-way multivariate analysis of variance results showed no significant differences between the test scores assigned to the two modes. Exploratory factor analysis did not reveal differences in the underlying factor structures of the two modes.

Conclusions: The findings provide evidence for the use of monologic tasks in computer-delivered speaking tests.

Keywords: Speaking test; Computer-delivered test; Face-to-face test; Interview test; Effects of delivery mode

Background

Recent improvements in computer technology have promoted computer delivery of speaking tests. The potential benefits of computer delivery are increased reliability of the test resulting from the standardization of delivery, more efficient test administration, faster score reporting, and the flexibility in the presentation of tasks with various sources of multi-media input.

The use of computers in speaking assessments is appealing; however, the absence of an interlocutor has resulted in concerns about the validity of using them as a replacement for interview tests. In test validation, language testers have long held an interest in specifying and minimizing the factors that confound score interpretation. For instance, mode effects, a facet of task conditions, have been discussed as a potential source of construct-irrelevant variance in computer-based tests. According to Chapelle and Douglas (2006), the most ubiquitous concern raised about assessing language via technology is that test takers' performance on a computer test may not reflect their ability measured by other forms of assessments. Therefore, there has been a call for

more research on the effects of computer-based testing (Alderson 2004); specifically, it appears that examinations comparing computer-based tests with conventional tests are needed (Chapelle 2003).

In addition, the validity of technology-based speaking tasks in eliciting linguistic performance has been questioned. For instance, studies of the task-based research have collected data using computer-delivered (e.g., Ellis 2005) or tape-based monologic tasks (Elder and Iwashita 2005; Iwashita et al. 2001; Wigglesworth 1997) in an effort to control for covariates in experiment designs. However, Elder and Iwashita (2005) suggested that test takers may not feel motivated to strive for better performance when an interlocutor is not present; this may result in inconclusive findings in such studies. Given that computer-delivered tasks also lack an interlocutor, this same validity issue may also apply to the computer mode.

Despite these concerns, few studies have investigated the effects of computer delivery mode on speaking tests; moreover, little is known about how the psychometric properties of computer-delivered speaking tests compare to face-to-face tests. For instance, previous studies have addressed issue of face validity (Kenyon and Malabonga 2001) and the effectiveness of technical aspects (Malabonga et al. 2005) of the Computerized Oral Proficiency Instrument. There are also studies that have examined test takers' strategic behaviors on the Speaking section of the Internet-based Test of English as a Foreign Language (Swain et al. 2009) and have compared test takers' performance on the test with their actual academic performance (Brooks and Swain 2014). In the few studies that have compared these two types of modes, the focus has been on test takers' speech samples (Zhou 2008) and the neural processes underlying performance on these modes (Jeong et al. 2011). For instance, Zhou (2008) found that test takers used more repetition words during the interviewer-delivered monologic tasks but more filled pauses during the computer-delivered monologic tasks. In addition, Jeong et al. (2011) concluded that direct interviews may elicit a more balanced and varied communicative ability than semi-direct interviews.

Given that there is an increase in the use of computers in speaking assessments, it is necessary to better understand how the mode of computer delivery affects speaking assessments. Therefore, in the present study, the psychometric qualities of computer-delivered and face-to-face monologic tasks were examined; specifically, test scores and the underlying factor structures of these tasks were compared. Indeed, professional testing standards (AERA et al. 1999) have underscored that score equivalence is significant and should be established prior to the interpretation of computerized test scores. Furthermore, exploring how the underlying factor structures vary between the two modes will provide valuable insight into the interpretation of test scores on computer-delivered speaking tests.

The comparability of the test scores of speaking tests between modes

Previous studies on the comparability of speaking test scores between modes have focused on the Oral Proficiency Interview (OPI) and the Simulated Oral Proficiency Interview (SOPI); these studies have shown evidence of score equivalence between the two tests. For instance, in a study of 10 individuals, Shohamy (2004) found no difference in mean scores between the Hebrew OPI and SOPI. Similarly, Kenyon and Tschirner (2000) revealed no difference between the scores of 20 students on the German OPI and SOPI. It is

important to note that the ratings in these studies were derived from the American Council on the Teaching of Foreign Languages scale, a holistic rating scale. Therefore, investigations of the differences between modes using analytic scales have not been conducted.

It is important to note that there were methodological limitations in the aforementioned studies. Specifically, the samples in both studies were small, thereby raising concerns about high sample dependency. It is also unclear if the same group of subjects was compared in the study by Shohamy (2004). Furthermore, Kenyon and Tschirner (2000) did not adopt a counterbalanced design; therefore, practice or fatigue effects may have occurred during the two tests. Finally, the OPI and the SOPI differ in task type and content; thus, it has not yet been determined if direct comparisons of these two tests are valid.

The comparability of the constructs underlying different modes of speaking tests

There is only one study that has compared the underlying constructs of different modes of speaking tests. O'Loughlin (2001) compared data from 83 test takers who took both the tape-based and live versions of the Australian Assessment of Communicative English Skills. Despite the high correlation between ability estimates (r = .92), the chi-square index (an assessment of the dimensionality of underlying constructs) was not statistically significant. Thus, a single dimension of speaking ability could not be constructed from the data combined from the tape and live versions. In addition, the lexical density estimates were significantly higher in the tape version than in the live version, thereby suggesting that the tape version elicited a more literate and formal type of language output. Therefore, O'Loughlin speculated that the live version may measure interactive ability, while the tape version may assess monologic ability. This is not surprising as the live version included a role-playing task that involved the test taker and the interviewer.

Therefore, it appears that the issue that requires further examination is the extent that monologic tasks contribute to the lack of unidimensionality between the two modes. The interviewers in the O'Loughlin (2001)'s study provided minimal responses; the interviewer-delivered monologic tasks should have measured the same ability as in the tape-based monologic tasks (i.e., monologic ability). However, lexical density estimate for the tape-based monologic tasks was significantly higher than the live version; this suggests that test takers used more formal language in the tape-based version. Therefore, it appears that even simple feedback provides a certain degree of interaction, compared to a condition where the interviewer is absent.

The co-construction of discourse during speaking tests may best explain this proposal. Performance during a direct test has been considered to be achieved jointly by participants in the interaction (Lazaraton 1996; McNamara 1997). Monologic tasks delivered by an interviewer who give minimal responses (e.g., backchannels) and non-verbal reactions (e.g., nodding and facial expressions) could also be considered co-constructed, however, with a partially responsive interactional partner. Therefore, interviewer-delivered monologic tasks may also measure interactive ability, but involve a lesser degree of interaction than interactive tasks (e.g., role-play).

O'Loughlin (2001) has made significant contributions to the understanding of the comparability of the two types of speaking tests; however, this study had two methodological limitations. First, while the study counterbalanced the testing orders, it did not separately analyze the scores of the groups assigned to the two different testing orders.

Indeed, an artifact of counterbalancing may be a differential carryover effect, which occurs "when the carryover effect of Treatment Condition 1 onto Treatment Condition 2 is different from the carryover effect of Treatment Condition 2 onto Treatment Condition 1" (Maxwell and Delaney 2004, p. 556). Thus, the potential differential carryover effects in the study could invalidate the analyses where the data from the two groups were combined. Second, the study paralleled tasks to minimize the task variable; however, no evidence of their supposed equivalence was offered. Thus, differences in task type and task content may be confounding factors that were not controlled for when measuring the mode effects.

The present study

The present study was designed to explore the effects of delivery mode (computer vs. face-to-face) on the psychometric aspects of speaking performance rated on analytic scales. A counterbalanced within-subjects design was adopted in order to eliminate the potential confounding factor of participant. Furthermore, differential carryover effects were checked to ensure that counterbalancing did not differently affect the two groups.

In addition, this study compared monologic tasks that have the same content between the two modes to address task as a potential confounding factor. As such, any differences observed were more likely to be attributed to mode effects than differences in other task-related variables. Furthermore, this type of design was implemented based on the fact that the current technology of automatic speech recognition technology cannot deliver a truly interactive task through computers. Therefore, a solution to this practical constraint was to focus on monologic tasks rather than compare an interview test that includes both monologic and interactive tasks with computer-delivered monologic tasks.

Specifically, this study addressed the following research questions:

1) Are there differences between delivery modes (computer vs. face-to-face) in the magnitude of the scores assigned to test takers' performance?
2) To what extent are the underlying factor structures of monologic tasks different or similar between delivery modes (computer vs. face-to-face)?

Methods

Participants

Participants included 79 Japanese students who learned English as a foreign language in Tokyo, Japan. There were 61 undergraduate students (77%) from three universities and 18 high school students (23%) from two high schools. Both university and high school students were recruited to obtain a sample representing a wide range of English proficiencies. The undergraduate students' major included foreign languages other than English (36%), English language and literature (13%), and domestic science (28%). The high school students were either from a boys' high school (13%) or a co-educational high school (10%). There were 19 males (24%) and 60 females (76%). Participants ranged in age from 17 to 22 (M = 19.7). Students reported five to ten years of previous English language instruction (M = 7.6) and participated on a voluntary basis.

Tasks

Two types of monologic tasks were used: a narrative task and an opinion task. Monologic tasks usually refer to tasks that elicit long individual discourses without test takers'

interacting with an interlocutor. These types of tasks include reading-aloud, sentence repetition, information transfer, and oral presentation (O'Sullivan 2008). As reading-aloud and sentence repetition tasks are not common during face-to-face tests, the current study focused on the information transfer and oral presentation tasks. In these tasks, test takers take some time to make several points and to develop an adequate reply to the task prompts. The narrative and opinion tasks represent the information transfer and oral presentation tasks, respectively.

The computer-delivered tasks used were a narrative task and an opinion task from the speaking section of the Global Test of English Communication for STUDENTS[a] (hereafter, GTEC for STUDENTS). The GTEC for STUDENTS was developed by the Benesse Corporation in Japan; this is a four-skill, computer-delivered English test that primarily targets Japanese high school and university students. The narrative task contained four pictures that told a simple story; participants had one minute to relate the pictures. A video prompt was also presented where an American female asked questions and gave simple preset feedback (e.g., "very good"). During the opinion task, a graph was provided; participants were required to give their opinions about a topic based on the graph within two minutes. The oral and written instructions were in Japanese. Participants were given two and three minutes to prepare their responses for the narrative and opinion tasks, respectively. During the preparation time, note taking was not allowed.

The face-to-face tasks were constructed using the same content and format as those delivered by computer; they were conducted on a one-to-one basis. A Chinese female (the author) who was proficient in both English and Japanese, served as the interviewer. Instructions for each task were written on a prompt card in Japanese. During the narrative task, the interviewer asked questions and gave feedback as the video character had done during the computer-delivered task. The same time periods allotted during the computer-delivered tasks for response and preparation time were provided.

Study design and procedure

Participants were randomly assigned to two groups in a counterbalanced design; specifically, there were two sessions. Group A ($n = 41$) completed the computer-delivered tasks during Session 1 and the face-to-face tasks during Session 2; Group B ($n = 38$) completed the face-to-face tasks during Session 1 and the computer-delivered tasks during Session 2. The two sessions had an interval of seven to ten days between them. Participants were not informed that the tasks would have the same content.

The computer-delivered tasks were administered either individually in a quiet room during school hours or in a computer lab during scheduled classes with approximately 20 classmates. In the computer lab, participants were seated far enough apart to ensure that they would not influence each other. Before the computer session, a video clip briefly demonstrated how to perform the tasks; this was done to familiarize the participants with the testing procedure. The participants started recording their responses by clicking a "start" icon on the screen. After the preparation time was over, responses were automatically recorded.

During the face-to-face session, participants met individually with the interviewer in a quiet room on their campus. Each participant was greeted by the interviewer in

Japanese and asked to sit at a desk facing her. For each task, the interviewer read the instructions in Japanese that were written on the prompt card. She timed their preparation and gestured to the participants when to begin their responses. The interviewer recorded participants' responses with a digital recorder. After the participants started responding to the tasks, the interviewer tried not to provide any verbal reaction except simple backchannels (e.g., mm-hm and uh-huh) with nodding and eye contact. Although the authenticity of the tasks may be questioned, the verbal reactions were controlled for three reasons. First, the narrative and opinion tasks were monologic; therefore, they required no interviewer input for completion. Second, limiting the potential variability in the interviewer's verbal reactions was desirable for issues related to task reliability. Finally, minimizing verbal reactions allowed for consistency with previous research (O'Loughlin 2001).

Scoring

Five accredited raters of the GTEC for STUDENTS scored the responses. The raters were native speakers of English who worked as English instructors in an English school in Tokyo. Two raters awarded ratings on an analytic scale of 1 to 4 for each of the four rating elements: grammar, vocabulary, fluency, and pronunciation. Participants may have received ratings from a different pair of raters; this is quite a common practice for large-scale performance tests (Lee 2006). Indeed, it is usually impractical to ask the same pair of raters to award all ratings on a task.

Participants' scores were determined by taking the mean of the two ratings. The same pool of raters scored the face-to-face tasks and used the same scoring rubric that was used for the computer-delivered tasks.

Analyses and results

Rating consistency

Two types of rater consistency indexes were calculated to assess the degree of consistency between the ratings: Pearson's correlation and rating agreement. The Pearson's correlation coefficients ranged from .52 to .75 for the computer-delivered tasks and .60 to .74 for the face-to-face tasks. The rating agreements for both modes were satisfactory with moderately high exact agreement (54.4%–68.4%) and adjacent agreement (29.1%–45.6%) for the face-to-face tasks; similarly moderate exact agreement (49.4%–72.2%) and adjacent agreement (26.6%–40.5%) were found for the computer-delivered tasks. The unsatisfactory correlation estimates may have been due to the fact that the rater pairs varied. When taking both types of indexes into account, the consistency of the ratings in both modes was considered acceptable herein.

Thus, the two ratings awarded to each element were averaged for each participant; the averaged ratings were referred to as the *element scores* in subsequent analyses. SPSS 12.0 for Windows was used for all the statistical analyses.

Test score comparison

Differential carryover effects

Prior to comparing scores, two repeated measures multivariate analysis of variance (MANOVA) were conducted at the task level to determine whether there were any significant mode-by-order interactions. For each MANOVA, delivery mode with two

levels was the within-subject factor; test order with two levels was the between-subjects factor. The dependent variables were the four rating elements. An alpha level of .05 was adopted for all analyses.

For the narrative task, the assumptions were first checked. A Box's M test was performed due to the unequal sample sizes, and the homogeneity of variance was confirmed. Levene's test of equality of error variances also indicated that the assumption of homogeneity of variance was not violated. The assumption of sphericity was tested by Mauchley's sphericity test. The test was significant ($p < .05$) across the four rating elements; therefore, the degrees of freedom were adjusted with the values of the Greenhouse-Geisser test in the follow-up univariate analyses of variance (ANOVA).

The descriptive statistics for the narrative task scores are presented in Table 1. Both Groups A and B had higher means for Session 2 for grammar, vocabulary, and fluency. This was expected due to practice effects. However, the differences in the mode means were larger for vocabulary (0.23) and fluency (0.19) than for grammar (0.05). In contrast, the mean pronunciation scores showed an opposite pattern: Group A had a higher mean for Session 2 (face-to-face: $M = 2.40$) than for Session 1 (computer: $M = 2.05$), whereas Group B had a higher mean for Session 2 (computer: $M = 2.05$) than for Session 1 (face-to-face: $M = 1.97$).

As shown in Table 2, there was a significant main effect for mode ($F(74, 4) = 5.91$, $p < .05$, $\eta^2 = .24$) and significant mode-by-order interaction effect ($F(74, 4) = 3.01$, $p < .05$, $\eta^2 = .14$). A significant simple main effect for mode was only found for pronunciation. However, mode-by-order interactions were significant for vocabulary ($F(77, 1) = 7.58$, $p < .05$, $\eta^2 = .09$), fluency ($F(77, 1) = 7.02$, $p < .05$, $\eta^2 = .08$), and pronunciation ($F(77, 1) = 6.59$, $p < .05$, $\eta^2 = .08$). The interaction effects (Eta squared η^2) were in the medium effect range as specified in Cohen (1977) ($\eta^2 = .01$, $\eta^2 = .06$, and $\eta^2 = .14$ as small, medium, and large effect sizes, respectively).

For the opinion task, homogeneity of variance was confirmed through Box's M test and Levene's test. However, the assumption of sphericity was violated according to

Table 1 Descriptive statistics of the narrative task scores

	Computer		Face-to-face		
Rating element	***M***	***SD***	***M***	***SD***	***n***
Grammar					
Group A (C → F)	2.05	0.81	2.15	0.74	41
Group B (F → C)	1.91	0.62	1.86	0.76	38
Vocabulary					
Group A (C → F)	2.21	0.77	2.51	0.79	41
Group B (F → C)	2.12	0.70	2.05	0.72	38
Fluency					
Group A (C → F)	2.01	0.79	2.10	0.78	41
Group B (F → C)	2.08	0.81	1.80	0.63	38
Pronunciation					
Group A (C → F)	2.05	0.68	2.40	0.57	41
Group B (F → C)	1.97	0.52	2.05	0.75	38

Note. C → F: Computer-delivered tasks first/face-to-face tasks second; Face-to-face tasks first/computer-delivered tasks second.

Table 2 Results of the repeated measures MANOVA for the narrative task

		df	*F*	*p*	η2
Overall effects					
Mode		4	5.91	.00	.24
Order		4	1.40	.24	.07
Mode × Order		4	3.01	.02	.14
Effects on different rating elements					
Mode	Grammar	1	0.16	.69	.00
	Vocabulary	1	3.15	.08	.04
	Fluency	1	1.96	.17	.03
	Pronunciation	1	16.35	.00	.18
Order	Grammar	1	1.90	.17	.02
	Vocabulary	1	3.17	.08	.04
	Fluency	1	0.53	.47	.01
	Pronunciation	1	2.55	.11	.03
Interaction effects on different rating elements					
Mode × Order	Grammar	1	1.80	.18	.02
	Vocabulary	1	7.58	.01	.09
	Fluency	1	7.02	.01	.08

Mauchley's test; therefore, the degrees of freedom were adjusted with the values of the Greenhouse-Geisser test in the follow-up univariate ANOVAs.

The descriptive statistics for the opinion task scores are displayed in Table 3. The mean fluency scores in Table 3 indicate that both Groups A and B performed better during Session 2, but the difference in the mode means was larger for Group A (0.33) than for Group B (0.12). For the other rating elements, Group A performed better during Session 2, whereas the reverse was true for Group B.

The results of the MANOVA shown in Table 4 revealed both a significant overall main effect for mode ($F(74, 4) = 11.66$, $p < .05$, $\eta^2 = .39$) and a significant mode-by-

Table 3 Descriptive statistics of the opinion task scores

	Computer		Face-to-face		
Rating element	***M***	***SD***	***M***	***SD***	***n***
Grammar					
Group A (C → F)	1.74	0.71	2.23	0.73	41
Group B (F → C)	1.75	0.71	1.87	0.72	38
Vocabulary					
Group A (C → F)	1.91	0.77	2.45	0.77	41
Group B (F → C)	1.84	0.71	2.03	0.75	38
Fluency					
Group A (C → F)	1.56	0.54	1.89	0.70	41
Group B (F → C)	1.71	0.74	1.59	0.63	38
Pronunciation					
Group A (C → F)	1.83	0.71	2.38	0.62	41
Group B (F → C)	1.71	0.65	1.92	0.69	38

Note. C → F: Computer-delivered tasks first/face-to-face tasks second; Face-to-face tasks first/computer-delivered tasks second.

Table 4 Results of the repeated measures MANOVA for the opinion task

		df	*F*	*p*	η2
Overall effects					
Mode		4	11.66	.00	.39
Order		4	2.26	.07	.11
Mode × Order		4	4.51	.00	.20
Effects on different rating elements					
Mode	Grammar	1	27.96	.00	.27
	Vocabulary	1	27.29	.00	.26
	Fluency	1	2.87	.09	.04
	Pronunciation	1	39.41	.00	.34
Order	Grammar	1	1.40	.24	.02
	Vocabulary	1	2.61	.11	.03
	Fluency	1	0.31	.58	.00
	Pronunciation	1	4.37	.04	.05
Interaction effects on different rating elements					
Mode × Order	Grammar	1	10.38	.00	.12
	Vocabulary	1	6.52	.01	.08
	Fluency	1	12.94	.00	.14
	Pronunciation	1	7.82	.01	.09

order interaction effect ($F(74, 4) = 4.51$, $p < .05$, $\eta^2 = .20$) for the opinion task. The univariate analysis yielded significant simple main mode effects for grammar, vocabulary, and fluency. The mode-by-order interaction effect was also significant for each rating element: grammar ($F(77, 1) = 10.38$, $p < .05$, $\eta^2 = .12$), vocabulary ($F(77, 1) = 6.52$, $p < .05$, $\eta^2 = .08$), fluency ($F(77, 1) = 12.94$, $p < .05$, $\eta^2 = .14$), and pronunciation ($F(77, 1) = 7.08$, $p < .05$, $\eta^2 = .09$). The interaction effects were in the range of medium and large effects as defined by Cohen (1977).

In sum, for both the narrative and opinion tasks, delivery mode yielded both significant overall and simple main effects, as well as significant mode-by-order interaction effects. The interpretation of the mode-by-order interaction effects should take precedence over the main effects.

One-way MANOVAs

Differential mode carryover effects were observed; therefore, following Vispoel et al. (2001), only data from Session 1 were used to compare scores between the two modes. Two one-way MANOVAs were performed separately for the narrative and opinion tasks, with delivery mode as the between-subjects variable and the four rating elements as the dependent variables. The assumptions of the MANOVA were confirmed through Box's M test and Levene's test. Given the significant results from Mauchley's test, the *p*-values associated with the *F* statistics were adjusted with the Greenhouse-Geisser correction.

As shown in Table 5, each mean element score on the narrative task was equal to or slightly higher for the computer mode; however, the score on the opinion task was slightly higher for the face-to-face mode. Nevertheless, the differences between the mode means were small for each rating element (less than or equal to 0.21). The results confirmed that for both the narrative and opinion tasks, there were no significant overall mode effects or

Table 5 Results of the one-way MANOVA for the narrative and opinion tasks

	Computer (*n* = 41)		Face-to-face (*n* = 38)					
	M	*SD*	*M*	*SD*	*df*	*F*	*p*	η2
Narrative task								
Overall effects					4	1.59	.19	.08
Grammar	2.05	0.81	1.86	0.76	1	1.19	.28	.02
Vocabulary	2.21	0.77	2.05	0.72	1	0.85	.36	.01
Fluency	2.01	0.79	1.80	0.63	1	1.67	.20	.02
Pronunciation	2.05	0.68	2.05	0.75	1	0.00	.98	.00
Opinion task								
Overall effects					4	0.25	.91	.01
Grammar	1.74	0.71	1.87	0.72	1	0.60	.44	.01
Vocabulary	1.91	0.77	2.03	0.75	1	0.43	.52	.01
Fluency	1.56	0.54	1.59	0.62	1	0.06	.81	.00
Pronunciation	1.83	0.71	1.92	0.69	1	0.34	.56	.00

simple mode effects for any rating element. The results showed that only minor differences between the mode means were observed; none reached significance.

Underlying factor structure comparison

An exploratory factor analysis (EFA) was performed on the element scores at the task level to investigate whether monologic tasks delivered via the two modes measured common components. The data were transformed by mean centering to address the differential mode carryover effects. This method involves subtracting the mean of each rating element from each case for Groups A and B, respectively, so that the new means are zero. Centering the data helps "remove the covariance associated with the different means while retaining that resulting from different deviations around the means" (Rummel 1988, p.292).

The transformed data from the two groups were then combined for the factor analysis. There were 16 variables—four element scores for each of the two tasks delivered via each of the two modes. The number of subjects should be at least five times the number of variables in the analysis (Field 2005); therefore, 79 data points for 16 variables was considered marginally acceptable.

The assumptions of EFA were checked. No variables deviated from normality, as the values for both skewness and kurtosis fell within an acceptable range of −2 to 2. Pearson's correlations between the variables ranged from .56 to .84, indicating that multicollinearity was not a problem. Finally, there were no univariate or multivariate outliers, as assessed by *z* scores and Mahalanobis distance, respectively.

A principal factor analysis of the 16 variables was conducted with varimax rotation. The solution produced single component on the basis of eigenvalues greater than 1 (see Table 6) and was confirmed by the scree plot. This single factor accounted for 71.12% of the total variance.

Table 7 shows the factor loadings of all of the variables. The variables loaded highly on the factor; loadings ranged from .78 to .88; each pair of variables generally had

Table 6 Exploratory factor analysis of combined data from the two modes

Factor	Eigenvalue	Percentage of variance	Cumulative percentage
1	11.38	71.12	71.12
2	0.87	5.41	76.53
3	0.79	4.97	81.50
4	0.61	3.79	85.29
5	0.44	2.78	88.07
6	0.41	2.55	90.62
7	0.32	2.03	92.65
8	0.22	1.40	94.05
9	0.19	1.17	95.22
10	0.17	1.07	96.29
11	0.15	0.92	97.21
12	0.13	0.83	8.04
13	0.10	0.62	98.65
14	0.08	0.52	99.17
15	0.07	0.42	99.60
16	0.06	0.40	100.00

equivalent loadings on the factor. Factor loadings reflect the portion of total variance each variable contributes to the factor; therefore, the results suggest that the variables at the task level contributed similarly to the major component when compared between modes. Overall, the EFA indicated that the monologic tasks delivered via the two modes contributed similarly to a unidimensional factor structure.

Table 7 Factor loadings of the 16 variables

Variable	Factor loading
Computer_Grammar_Narrative	.86
Computer_Vocabulary_Narrative	.85
Computer_Fluency_Narrative	.82
Computer_Pronunciation_Narrative	.85
Computer_Grammar_Opinion	.85
Computer_Vocabulary_Opinion	.85
Computer_Fluency_Opinion	.78
Computer_Pronunciation_Opinion	.87
Interview_Grammar_Narrative	.86
Interview_Vocabulary_Narrative	.86
Interview_Fluency_Narrative	.83
Interview_Pronunciation_Narrative	.85
Interview_Grammar_Opinion	.88
Interview_Vocabulary_Opinion	.85
Interview_Fluency_Opinion	.81
Interview_Pronunciation_Opinion	.83

Discussion

The comparability of test scores

The results of the MANOVA from Session 1 data did not revealed a difference between the mode means for each rating element. Thus, at the task level, there seems to be no mode effect on the test scores assigned to each analytic scale.

One interpretation of this set of results is that the participants performed similarly between the modes; this led the raters to assign similar ratings to their performance during each of the modes. If that was the case, the findings for the analytic scales were consistent with those of studies on tape-based speaking tests that used the ACTFL holistic scale (Kenyon and Tschirner 2000; Shohamy 2004). Thus, it appears that delivery mode is not a factor that affects the magnitude of scores, regardless of the type of rating scale used. However, more empirical evidence is needed to support this conclusion.

It may also be that the participants performed differently between the two modes, but not to an extent that the raters were able to discern. Indeed, the lack of score sensitivity to task variation is well documented in the language testing research. In a review of the literature, Fulcher (2003) concluded that studies that have found significant differences in performance across tasks used maximally different tasks (Chalhoub-Deville 1995) or multiple rating scales (Upshur and Turner 1999). In the present study, the tasks differed only in delivery mode, and the rating scales were the same for the two modes. If delivery mode is considered one dimension of the speaking task, the unobserved differences in test scores is not surprising.

It may also be possible that the participants responded to the tasks differently between the modes; however, the raters may have adjusted the possible variability in performance as reported in Brown (2005) and McNamara and Lumley (1997). Some raters in the current study may have awarded a higher rating to the participants, assuming that the interviewer was not helpful in eliciting speech samples. Raters may also have given a higher rating to the participants who were perceived as being disadvantaged by talking to a computer. Unfortunately, the data collected herein are insufficient to evaluate these hypotheses.

Finally, the lack of the differences in test scores between the modes may be attributed to the limited range of proficiency of the participants. Test takers with insufficient proficiency may perform better with an interviewer as they may feel reassured when the interviewer nods and gives minimal responses. In contrast, higher-proficiency test takers may not be influenced by the absence of an interlocutor. This study attempted to recruit high school students to represent the less proficient test takers. However, since they were volunteers, these students' teachers indicated that they had higher levels of oral proficiency than their classmates.

Although no mode effects were detected, mode-by-task type interactions may have existed. As seen in Table 5, during the narrative task, the participants performed better in the computer mode, whereas during the opinion task, they performed better in the face-to-face mode. However, it is difficult to reach a firm conclusion regarding the opposite trends given that there were several potential confounding factors. First, due to practical constraints, this study did not counterbalance the order of the narrative and opinion tasks; therefore, task order may have confounded any mode-by-task type interactions. Second, the narrative and opinion tasks differed in task difficulty and task

topic; this makes it difficult to attribute the observed opposite trends exclusively to mode-by-task type interactions.

Differential carryover effects were found, despite the fact that there was an interval of seven to ten days between the two sessions. This may be due to practice effects induced by the two delivery modes. The participants who took the computer-delivered tasks first may not have achieved the best performance because of the unfamiliarity of the task format. Thus, when given the second opportunity, in addition to the familiarity of the task content, the reactions from the interviewer may have motivated them to give better verbal responses. In contrast, those who completed the face-to-face tasks first may have hoped to achieve a higher score on the subsequent computer-delivered tasks, but may not have strived for better performance since there was no reaction from the computer. Therefore, the first administration of the face-to-face tasks was not beneficial for performance on the computer-delivered tasks.

A discussion of the different functions of the two modes regarding practice effects has implications for the design of comparability studies. Indeed, when there is reason to expect a differential mode carryover effect, researchers implementing comparison studies with a counterbalanced design are strongly encouraged to check for order effects and potential mode-by-order interaction effects. If such interaction effects are present, the original experiment should be reanalyzed as a between-subjects design by examining data only from the initial session. However, this would lead to a reduction in data and may affect the statistical power of the analysis. Therefore, it is essential to conduct such studies with large samples. Moreover, when a large sample size is not feasible due to practical constraints, it may be necessary to abandon the within-subjects design and adopt a between-subjects design.

The comparability of the underlying factor structures of the two modes

The results of the EFA revealed a single factor for the combined data from the two modes. In addition, all the variables loaded highly on the single factor, and their factor loadings generally showed a similar pattern between the modes. These findings imply that monologic tasks delivered via the two modes measure a similar underlying factor.

However, these findings were not congruent with the prediction that technology-based monologic tasks measure a different underlying construct from interviewer-delivered ones. These unexpected results may be related to the fact that this study used exploratory factor analysis. Future studies should consider using more sophisticated statistical procedures, such as multi-group confirmatory factor analysis. While this approach requires a larger sample, it would provide more information about the potential differences in factor loadings and underlying factor structures; thus, it allows for a more rigorous comparison of the psychometric properties of monologic tasks between modes.

In addition, it is important to note that this study examined mode effects on the constructs from a psychometric perspective; therefore, the findings did not address what kind of speaking ability this single factor may represent. Zhou (2008) reported the finding on the lexical density from analyzing speech sample; this study used the same participant group that was used herein. Thus, the findings from this study may shed some light on this question. The lexical density estimates reported by Zhou showed an insignificant difference between the computer mode (62%) and face-to-face mode

(63%); this supports the current finding that there is a common factor between performances in the two modes.

Furthermore, compared to the literature extant, the magnitude of the lexical density in Zhou (2008) was quite high in both modes. For example, Ure (1971) found that spoken texts had a lexical density less than 40% and written texts had a lexical density greater than 40%. Therefore, the speech samples elicited in the present study could be considered similar to written language or monologic discourse, which represents a monologic speaking ability. In addition, the high figures on lexical density suggest a similarly low degree of interaction is involved in monologic tasks delivered via the two modes. However, although lexical density provides important clues about the speaking abilities measured by monologic tasks, it does not actually assess any meaningful component of language proficiency, as noted by O'Loughlin (2001). Therefore, the interpretation of the underlying single factor discussed herein is tentative.

Conclusions

This study examined differences in performance on computer-delivered monologic tasks and face-to-face monologic tasks. The results provided evidence for the validity of computer-delivered monologic tasks. The fact that there were no significant differences observed in test scores implies that the results on computer-delivered monologic tasks could be used to infer scores on face-to-face monologic tasks. Moreover, the same underlying factor structures measured by monologic tasks in the two modes suggest that scores on computer-delivered monologic tasks could be interpreted similarly to those scores on face-to-face monologic tasks.

One possible limitation of this study is that the implications stated above are only applicable to (a) Japanese learners, (b) two monologic tasks, and (c) tasks with a female interviewer. More specifically, the participants in this study were Japanese learners in Japan; therefore, the results cannot be generalized to other nationalities. In addition, a narrative task and an opinion task were used herein; the results can only be interpreted in relation to these two particular types of monologic tasks. Furthermore, the relationship between monologic ability and interactive ability is still unclear (Iwashita et al. 2008); thus, generalizing the current results to interactive tasks should be done with caution. Moreover, a female delivered all of the face-to-face tasks. Research has found that women feel more comfortable talking to a female audience during an interview (O'Sullivan 2000). Since two-thirds of the participants in the current study were female, if a male had conducted the interviews, the results may have differed.

In addition, the current study had methodological limitations. For instance, the proficiency of participants assigned to the different testing conditions was not known to be equivalent; this was a concern. A within-subjects design was originally planned to compare test scores however, a between-subjects analysis was conducted when practice effects were observed. In these cases, it is critical that the two groups are equivalent to support the internal validity of the findings. Although the participants were randomly assigned to the two groups, there was no objective index assessing the equivalence of the participants' proficiency between the two groups.

This study was also limited by exclusively focusing on psychometric properties. There is a need to gather comprehensive evidence to support the notion that the two

delivery modes are equivalent; this would provide additional evidence of the validity of computer-delivered monologic tasks (Weir 2005). Therefore, it is recommended that future research compare test takers' speech samples, their use of strategies, and their perceptions of the two modes. Furthermore, it is essential to explore how individual characteristics of test takers, such as computer anxiety and computer familiarity, may mediate the relationship between delivery mode and the testing of speaking skills; this is critical to establishing whether or not computer-delivered speaking tests are fair.

Finally, future studies should focus on task prompts and test takers' actual use of planning time in order to advance the understanding of the effects of computer delivery mode on speaking tests. Today, the multi-media capability of computers has made both audio and video forms available. It would be interesting to investigate how test takers' performance differs according to these types of feedback. Moreover, it is important to note that most of the participants did not rehearse their speech aloud during the face-to-face tasks; however, during the computer-delivered tasks, they may have combined preparing for the tasks with rehearsing for the tasks. Thus, future research should examine test takers' different approaches for planning by videotaping their behavior and conducting retrospective interviews.

Endnote

[a] The speaking section of the GTEC for STUDENTS used in this study is not in use as of 2009.

Competing interests
The authors declare that they have no competing interests.

Acknowledgments
This article is based on the author's doctoral dissertation submitted to Tokyo University of Foreign Studies. I am grateful to Masashi Negishi for his guidance, Masanori Ichikawa for his constructive advice on data analysis, and Hideyuki Takashima, Yukio Tono, Tae Umino, and Yuko Nakahama for their insightful comments. I thank Benesse Corporation for giving permission to use the speaking part of the GTEC for STUDENTS and for their cooperation in rating the test. Finally, thanks go to the anonymous reviewers for their advice in revising the article.

Received: 26 September 2014 Accepted: 15 December 2014
Published online: 03 February 2015

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, C. (2004). The shape of things to come: will it be the normal distribution? In M Milanovic & C Weir (Eds.), *Studies in language testing 18: European language testing in a global context* (pp. 1–26). Cambridge: Cambridge University Press.

Brooks, L, & Swain, M. (2014). Contextualizing performances: comparing performances during TOEFL iBT and real-life academic speaking activities. *Lang Assess Quart, 11*(4), 353–373. doi: 10.1080/15434303.2014.947532.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt am Main: Peter Lang.

Chalhoub-Deville, M. (1995). A contexualized approach to describing oral language proficiency. *Lang Learn, 45*(2), 251–281. doi: 10.1111/j.1467-1770.1995.tb00440.x.

Chapelle, CA. (2003). *English language learning and technology*. Amsterdam: John Benjamins.

Chapelle, CA, & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Elder, C, & Iwashita, N. (2005). Planning in language testing. In R Ellis (Ed.), *Planning and task performance in a second language* (pp. 217–238). Amsterdam: John Benjamins.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: a psychometric study. *Stud Sec Lang Acquis, 27*(2), 141–172. doi: 10.1017/S0272263105050096.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman/Pearson Education.

Iwashita, N, Brown, A, McNamara, T, & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Appl Linguist, 29*(1), 24–49. doi: 10.1093/applin/amm017.

Iwashita, N, McNamara, T, & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Lang Learn, 51*(3), 401–436. doi: 10.1111/0023-8333.00160.

Jeong, H, Hashizume, H, Sugiura, M, Sassa, Y, Yokoyama, S, Shiozaki, S, et al. (2011). Testing second language oral proficiency in direct and semidirect settings: a social-cognitive neuroscience perspective. *Lang Learn, 61*(3), 675–699. doi: 10.1111/j.1467-9922.2011.00635.x.
Kenyon, DM, & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Lang Learn Technol, 5*(2), 60–83.
Kenyon, DM, & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: comparing performance at lower proficiency levels. *Mod Lang J, 84*(1), 85–101. doi: 10.1111/0026-7902.00054.
Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Lang Test, 13*(2), 149–170. doi: 10.1177/026553229601300202.
Lee, YW. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Lang Test, 23*(2), 131–166. doi: 10.1191/0265532206lt325oa.
Malabonga, V, Kenyon, DM, & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Lang Test, 22*(1), 59–92. doi: 10.1191/0265532205lt297oa.
Maxwell, SE, & Delaney, HD. (2004). *Designing experiments and analyzing data: a model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
McNamara, TF. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446–466.
McNamara TF, Lumley T. The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. Lang Test. 1997;14(2):140–56. doi: 10.1177/026553229701400202.
O'Loughlin, K. (2001). *Studies in language testing 13: the equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.
O'Sullivan, B. (2000). Exploring Gender and Oral Proficiency Interview Performance. *System, 28*(3), 373–386.
O'Sullivan, B. (2008). *Modelling performance in tests of spoken language*. Frankfurt, Germany: Peter Lang.
Rummel, RJ. (1988). *Applied factor analysis*. Evanston, IL: Northwestern University Press.
Shohamy, E. (2004). The validity of direct versus semi-direct oral tests. *Lang Test, 11*(2), 99–123. doi: 10.1177/026553229401100202.
Swain, M, Huang, LS, Barkaoui, K, Brooks, L, & Lapkin, S. (2009). *The Speaking Section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors* ((TOEFL iBT Research Rep. No. 10)). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-09-30.pdf. Accessed 21 Nov 2014.
Upshur, JA, & Turner, CE. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Lang Test, 16*(1), 82–111. doi: 10.1177/026553229901600105.
Ure, J. (1971). Lexical density and register differentiation. In GE Perren & JLM Trim (Eds.), *Applications of Linguistics* (pp. 443–452). Cambridge: Cambridge University Press.
Vispoel, WP, Boo, J, & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg self-esteem scale: a comparison of psychometric features and respondent preferences. *Educ Psychol Meas, 61*(3), 461–474. doi: 10.1177/00131640121971329.
Weir, CJ. (2005). *Language testing and validation: an evidence-based approach*. Basingstoke: Palgrave Macmillan.
Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Lang Test, 14*(1), 85–106. doi: 10.1177/026553229701400105.
Zhou, YJ. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *Jpn Lang Test Assoc J, 11*, 189–208. http://ci.nii.ac.jp/els/110009607589.pdf?id=ART0010071097&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1421407388&cp=. Accessed 20 Nov 2014.