THE JOURNAL OF ASIA TEFL
Vol. 14, No. 4, Winter 2017, 717-735
http://dx.doi.org/10.18823/asiatefl.2017.14.4.9.717

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 © 2004 AsiaTEFL.org. All rights reserved.

Testing the Comparability of Different L2 Oral Test Tasks

Siwon Park
Kanda University of International Studies, Japan

This study examined the comparability of three L2 oral test tasks (topic discussion, information gap, and semi-direct speaking) in the measurement of the four oral traits (pronunciation, fluency, grammar, and vocabulary). In examining the comparability, a CFA approach to a correlated traits/correlated methods design was employed, and evidence of convergent and discriminant validity was tested in the measurement of the traits by the methods. Through the model comparison approach (Byrne, 2006; Wideman, 1985), the hypothesized model was compared against a series of more restrictive models using χ^2 difference tests, and the information from the path coefficients was examined to check the equivalence further. The findings revealed that the method effect in the measurement model was so strong that the trait was not fully reflected in the measurement process. Also, the effect of method on trait was not consistent in that topic discussion produced most trait-related information among the test tasks. Moreover, methods were found discriminant although traits were not. Thus, the three methods cannot be considered comparable in measuring the L2 oral traits, and such incomparability was particularly evident between the two methods of group oral and semi-direct speaking.

Keywords: L2 oral test tasks, method effect, CTCM CFA, group oral, semi-direct speaking

Introduction

The use of tasks in assessing second language (L2) learners' oral proficiency has become popular in the field of L2 teaching and testing, and their influence on examinee performance has been explored and documented through numerous empirical studies. However, as Skehan (2001) and Fulcher (2003) noted earlier, the results of these studies were often not conclusive as to the effects of tasks on L2 production. In these studies, task performance was often defined too narrowly and, as a consequence, its variance couldn't be fully identified and investigated. Also, the tasks examined in these studies were not characteristically distinct and did not help account for their differential effects on examinee performance.

Tasks can be defined by their levels and studied as such. They can be studied at the levels of behavior, speech act, or structure. At the structural level, tasks can be characterized according to their features such as goal orientation (±convergent) and interaction (±required). Such characterization of tasks can help examine the extent to which they simulate qualitatively and quantitatively different speech performance by L2 learners (Swain, 2001).

The current study also concerns L2 oral tasks at the structural level and aims to examine the construct equivalence across three types of L2 oral test tasks: topic discussion, information gap, and semi-direct speaking. While these tasks have gained popularity in L2 teaching and assessment, they are also known as requiring the examinees to operate and apply different cognitive and linguistic skills in their speech performance. Such a differential effect of tasks will significantly limit the inference about an examinee's speaking ability especially if it is drawn based on a single task performance. Therefore, more research endeavors are called for to promote an understanding of how L2 oral (test) tasks with their unique characteristics affect examinee performance (Brown, 2003), which is the primary purpose of this study.

Prior Studies on Assessing L2 Oral Proficiency

Interests in oral proficiency in L2 testing have generated numerous studies on the use of direct (or interview), semi-direct, and/or group discussion tests. Many of them concern the use of interview tests for L2 speaking such as oral proficiency interviews (OPIs) and also that of semi-direct speaking tests delivered on a computer such as TOEFL *i*BT. There is also a line of research that examined the comparability of the same oral tests in differing formats (e.g., OPI vs. computerized OPI). Also, group oral tests have become popular in recent years due to their nature that resembles the authentic communication in real life situations. However, empirical studies that cross-examined two or more of these tests in light of their comparability are scarce.

Early studies of speaking tests mostly concern the issue of construct validity. Bachman and Palmer (1981) and Henning (1983) explored the construct validity of the FSI oral interview. Bachman and Palmer (1981) employed the Campbell-Fiske multi-trait multi-method matrix and confirmatory factor analysis to examine the construct validity of the interview test with two traits of speaking and reading and three methods of interview, translation, and self-rating. The results confirmed the presence of both convergent and discriminant validity of the FSI interview. In contrast, Henning (1983) examined the comparability of interview, imitation, and completion methods in assessing L2 oral proficiency. He found that the presence of validity evidence was not consistently observed across the three testing methods, and they were related to different trait components to differing degrees.

While Bachman and Palmer (1981) and Henning (1983) mainly concerned the construct equivalence of different oral proficiency tests, Duff (1993) and Weir and Wu (2006) studied different types of semi-direct test tasks and examined their impact on L2 oral performance. Duff (1993), based on her longitudinal observation of an L2 learner’s interlanguage development, examined the differential effects of discussion, picture description, and storytelling test tasks. However, her results were not conclusive, leading her to suggest adopting tasks that are more distinct for any future studies of relevant topics. Similarly, Weir and Wu (2006) examined the equivalence of three test forms, each of which included read-aloud, answering questions, and picture description test tasks. The results confirmed that the three test forms were statistically parallel, and two of them were parallel even at the individual task level.

Other researchers examined the comparability of the direct and indirect versions of the same speaking tests. Luoma (1997), for example, compared the direct and indirect versions of the Finnish National Certificate in Language Proficiency. The results showed that both versions were highly correlated, with the indirect version containing more situations and helping elicit more functions. Despite the high correlation between the two versions, however, the discourse produced by them was not comparable. Similarly, O’Loughlin (2001), using diverse analytical techniques, looked at the direct and indirect forms of *access*, a test developed to screen immigrants to Australia. His results, unlike Luoma’s, revealed that the two forms were not equivalent, and the discourse produced on each form varied because most likely the examinees applied different test-taking strategies and processes to each form.

Overall, the above-mentioned studies on speaking tests of different forms or formats are mostly concerned with their construct equivalence. Most of these studies inform close relationships between and among the tests in differing formats, though some of them also demonstrate differences with their interactive and discourse features and test-taking strategies.

Another line of studies on L2 oral assessment concerns the use of the group oral format. Roughly, three groups of researchers examined the use of group oral format in assessing L2 oral proficiency, each with a focus on the different characteristics of the test format.

One group of studies addresses the logistical aspects of the group oral format over other oral proficiency tests (e.g., Bonk & Ockey, 2003; Folland & Robertson, 1976; Fulcher, 1996, 2003; Hildon, 1991). They argue that the group oral format can lessen the examinee anxiety and therefore generate a natural discourse. They also point out that the test format leads to positive washback due to their interactive nature of oral communication while helping deal with logistic issues in test administration.

Another group explored the validity aspects of the group oral format as a method to assess L2 oral proficiency (Bonk & Ockey, 2003; Fulcher, 1996; Leaper & Riazi, 2014; Shohamy, Reves, & Bejarano, 1986; O'Sullivan, 2002; Park, 2008; Van Moere, 2006). Fulcher (1996) argued that the group discussion test assessed the same language skills as other test tasks such as the picture-based discussion and text-based discussion. Bonk and Ockey (2003), in their FACETS study, proposed group oral testing as a reliable and pedagogic technique as it closely resembles L2 classroom practices. However, Shohamy et al. (1986) found that the group oral task was related the least with other oral tasks such as discussion, reporting, and role play. Van Moere (2006) also noticed that test occasion and interlocutor characteristics were the factors that significantly affected performance variance and suggested that a group oral test may not be suitable for high-stakes testing. Likewise, Leaper and Riazi (2013) found that examinee performance in their group oral exams considerably varied depending on which prompts they were assigned, the implication of which is similar to that of Van Moere (2006).

The last group of studies on group oral testing deals with the validity issue, examining the group oral format from the perspectives of interlocutor characteristics and test discourse (Berry, 2004; Gan, 2010; Ockey, 2009; O'Sullivan, 2002; Nakatsuhara, 2011, 2013). They argue that the examinees' personality types affect their performance to the degree that is not negligible. As O'Sullivan (2002) contends, any test format with interactive features must carefully reflect the effects of personality type (e.g., introversion vs. extroversion), acquaintanceship and the number of interlocutors (three vs. four), and proficiency levels (lower vs. upper). They significantly and negatively in most cases affect the elicitation of substantive interactions and conversations and, eventually, the examinee performance.

As reviewed thus far, numerous studies have investigated the various facets of direct, semi-direct, and/or group discussion in L2 oral proficiency; nevertheless, these test tasks were rarely cross-examined with each other. In this study, therefore, three test tasks of topic discussion, information gap, and semi-direct tasks were employed and cross-examined with regards to their statistical comparability. In short, in order to answer the research question, *to what extent the three tasks are comparable in assessing L2 learners' oral proficiency*, the convergent, as well as the discriminant validity of the test method and trait were examined using a CFA approach to a correlated traits/correlated methods (CTCM) design.

Method

Participants

A total of 187 Japanese learners of English participated in this study. They were all English majors at a university in Japan, and their class standing varied as 47 juniors, 79 sophomores, and 61 freshmen. Out of the 187 students, 146 were female (78%), and the rest were male (22%).

For scoring the examinee performance, 10 experienced EFL teachers served as raters. They were full-time instructors at the English Language Institute of the university where this study took place. They were all experienced in assessing Japanese students' English oral performance as they had already served as raters multiple times for in-house speaking tests. All of them were native speakers of English with a postgraduate degree in applied linguistics or related. Seven of them were male and the rest female; their nationalities varied – five American, two British, two Australian, and one Singaporean.

Test Instruments and Rating Scales

Three test tasks were employed for the examination of their comparability: topic discussion, information gap, and semi-direct speaking tasks. Table 1 highlights the different characteristics of the test tasks at their structural level. Seven different features were identified with each of the tasks which help manifest their unique characterizations. Such task-specific aspects of the three tasks are expected to offer means to explore the possibility of their differential effects on L2 examinees' oral performance.

TABLE 1
Task Type and Their Characteristics (adapted from Pica, Kanagy, and Falodun, 1993)

		Group oral tasks		Semi-direct (e.g., picture description)
		Topic discussion	Information gap (Jigsaw)	
1	Information Holder	X = Y	X or Y	X
2	Information Requester	X = Y	Y or X	none
3	Information Supplier	X = Y	X or Y	X
4	Information Requester-supplier relationship	2 way>1 way (X to Y & Y to X)	2 way (X to Y/Y to X)	none
5	Interaction requirement	-required	+required	-required
6	Goal orientation	-convergent	+convergent	-convergent
7	Outcome options	1+/-	1	unspecified

As Table 1 exhibits, topic discussion is a typical open task with a feature of *-convergent* and with an *unspecified* task goal. The interlocutors are required to deliver their ideas and opinions on a given topic; yet, there are no explicit task goals for all interlocutors to achieve in concert. On the other hand, the information gap task is goal-oriented with a feature of *+convergent*. The interlocutors are obliged to work cooperatively to achieve the explicit task goal by sharing the information in hand.

In this study, four examinees (or three) sat together in a group and completed both topic discussion and information gap tasks. For the discussion task, each group was assigned a prompt and told to discuss the topic in the prompt with each other. For the information gap task, each group was assigned a jigsaw task, named *A Souvenir Task* and was asked to decide on a souvenir for their friend living in the U.S. whom they were going to visit soon. Each examinee in the group was provided different information about their friend and also suggestions for a possible souvenir item. Examinees in the group had to decide the best souvenir item while exchanging the information that they had in hand. For both discussion and information gap tasks, examinees were assigned 2 minutes for preparation and 10 minutes for the actual group discussion. The semi-direct speaking test involved no interlocutor, and instead, the recorded stimuli were used to elicit examinee responses. The test was delivered on a computer just like TOEFL *i*BT or the speaking version of the TOEIC test, and their responses were audio-recorded and also videotaped. Table 2 summarizes the specifications of the three semi-direct speaking tasks. The test consisted of four test tasks, the first of which was for warm-up. The examinees were given planning time and also response time for each task. When the time limit expired, the test screen automatically changed for the next task.

TABLE 2
Specifications of the Three Semi-direct Speaking Tasks

Task title	# of items	Planning time	Response time
Warm-up task	5		80 sec
Picture task	1	60 sec	120 sec
Map task	3	30 sec	150 sec
Speech task	1	60 sec	180 sec

In order to assess examinees' oral performance on the three tasks, raters used a rating scale with four oral traits of pronunciation, fluency, grammar, and vocabulary and five score bands from 0 to 4.0 with a half-point scoring assignment in each band which makes a total of nine possible scores (see Appendix). The rating scale was developed by the local university in Japan where the current study took place. The scale has been used for the purpose of assessing students' oral proficiency at the university, and a number of studies have been conducted in order to ensure its quality (e.g., Bonk & Ockey, 2003; O'Donnell & Park, 2008).

Procedure

Two training sessions were held for the raters before the tests were administered to the examinees. The first training session was for the concurrent rating of the two group oral tests and the other for the ratings of the recorded speech responses from the semi-direct tests. Over the two training sessions, the raters familiarized themselves with the specifications of the scales.

During the administration of the tests, the test tasks were presented counterbalanced to the examinees so as to control for the order effects. Test administration lasted for 3 days, and the scoring of the recorded responses 2 additional days.

All examinee responses on the three test tasks were double-rated. Each rater was assigned to different pairs after completing several sessions so that all the scores could be linked and subjected to FACETS as fully connected and the fair average scores could be generated. The scores from the three individual semi-direct tests were all averaged out for each examinee and were subjected to the analyses as such.

Results

Out of 187 score sets, 29 with missing values were all deleted listwise, and the rest of 158 fair-average score sets were entered into the analysis. Table 3 presents the descriptive statistics of the data.

TABLE 3
Descriptive Statistics of Univariate Test Variables (N = 158)

Variable label	*M*	*SD*	Min	Max	$Z_{skewness}$	$Z_{kurtosis}$
Topic discussion						
V1: Pronunciation	6.15	0.97	4.00	8.88	1.59	0.14
V2: Fluency	6.19	1.12	3.00	8.82	-0.04	-0.02
V3: Grammar	5.87	0.94	4.00	8.81	2.37	0.60
V4: Vocabulary	6.03	1.01	3.50	8.78	1.04	-0.63
Information gap						
V5: Pronunciation	6.09	1.00	3.50	8.90	-0.02	0.13
V6: Fluency	6.27	1.10	2.50	8.84	-1.03	1.37
V7: Grammar	5.85	0.94	4.00	8.80	1.89	-0.02
V8: Vocabulary	5.80	0.97	3.50	8.92	2.28	1.17
Semi-direct (Picture)						
V9: Pronunciation	5.78	1.13	2.67	8.84	-0.10	-0.05
V10: Fluency	5.56	0.89	3.00	7.63	0.38	-0.40
V11: Grammar	5.70	0.93	2.98	8.15	-0.54	0.44
V12: Vocabulary	6.15	0.97	4.00	8.88	1.59	0.14

Note. Mardia's coefficient (G2,P) = 32.62, Normalized estimate = 7.17.

Overall, the mean scores are similar; most of them range from 5.56 with SD of 0.89 (fluency of semi-direct) to 6.19 with SD of 1.12 (fluency of topic discussion). Compared to the mean scores of the traits under the methods of topic discussion and information gap, those of the traits under the semi-direct methods are relatively lower. As for the normality indices of skewness and kurtosis, two variables indeed resulted in *Z*-skewness beyond the acceptable range of ±2 (grammar of topic discussion = 2.37; vocabulary of information gap = 2.28). Therefore, *robust* maximum likelihood (RML) was applied to analyzing the test data, rather than a regular maximum likelihood method that Hu and Bentler (1999) recommended. EQS 6.1 for Windows (Bentler, 2003) was employed for the subsequent CTCM analyses.

In the general CFA approach to CTCM analyses, Widaman's (1985) guidelines are employed to test for the evidence of construct validity. The hypothesized CTCM model is compared against an alternatively nested series of more restrictive models (Byrne, 1994). The difference in χ^2 (i.e., $\Delta\chi^2$) between the hypothesized model and each nested model serves as the evaluation criterion for the evidence of construct validity.

This model comparison approach is adopted as it helps examine whether methods (i.e., test tasks) are responsible for the differential effects of the test tasks on examinee performance. Also, the comparisons of the individual parameters of the same traits and methods across different tasks further help examine the comparability of the tasks in the measurement of the traits. The hypothesized model is shown in Figure 1.

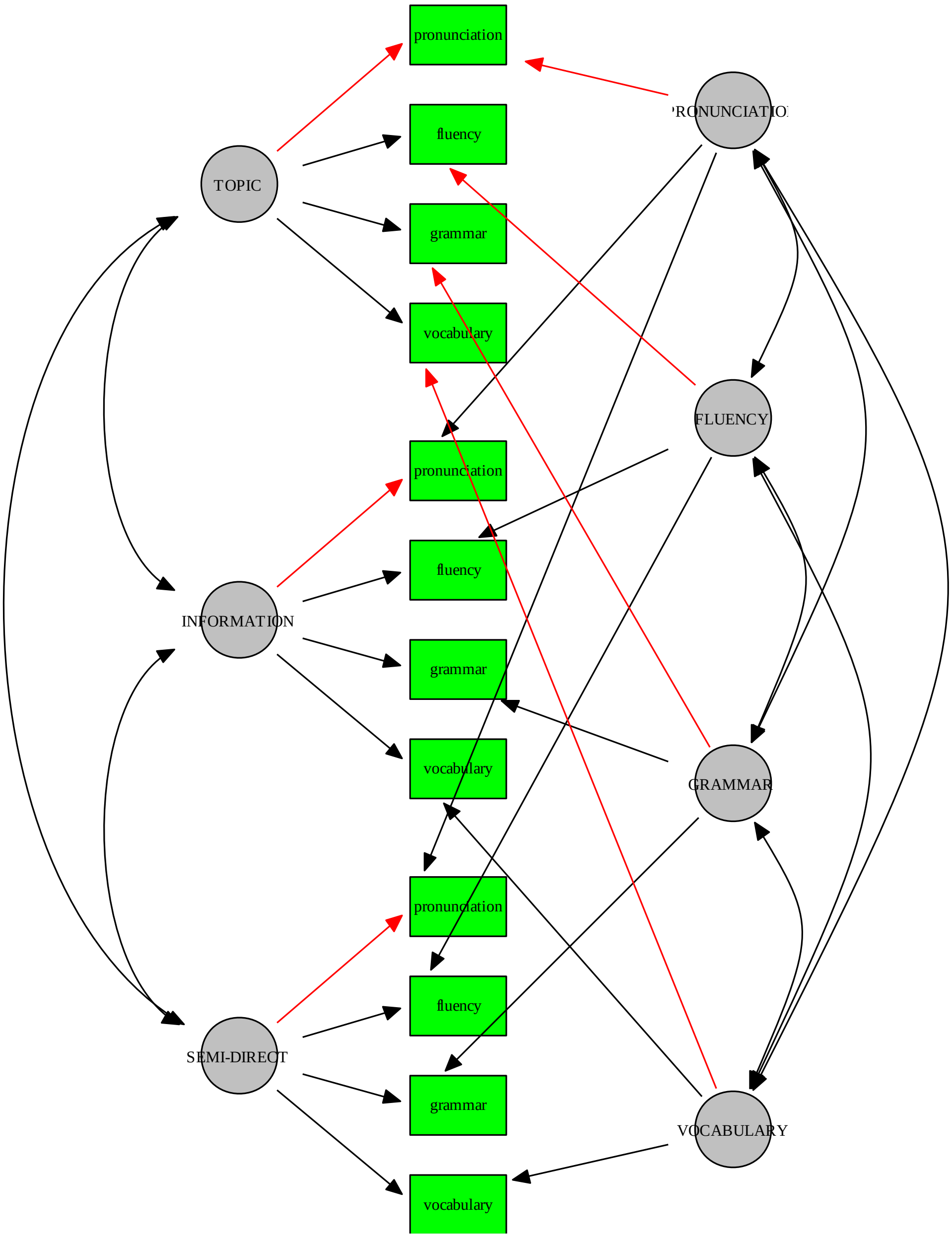

Figure 1. The hypothesized CTCM model with three methods and four traits (Model 1).

In the hypothesized model in Figure 1, four facets of oral proficiency – pronunciation, fluency, grammar, and vocabulary – are postulated on the right-hand side as measured by the three methods of topic discussion, information gap, and semi-direct speaking on the left-hand side. The squares in the middle denote the measurement variables of each trait in the test. Therefore, the model portrays a three by four method-trait CFA design with 12 measurement variables. The arrows specify the factor loadings, signifying the impact of the trait on each measurement variable. Although not observable in Figure 1, measurement variables are estimated together with their error terms.

Comparisons of the Hypothesized and Alternative Models

In the subsequent CTCM CFA analysis, models are compared pairwise and the results are interpreted, following Byrne's (2006) guidelines. For the model comparison, the hypothesized model of *freely correlated traits and freely correlated methods* in Figure 1 was first analyzed for its goodness-of-fit results. This hypothesized model is the least restrictive and serves as the baseline model against which all alternative models are compared.

Alternative models were estimated to obtain parameter estimates one after another. The first alternative was a model of *no traits and freely correlated methods*, as portrayed in Figure 2 (Model 2). The model assumes that there are no traits, but methods exist. A good model-data fit with this methods-only model will confirm the assumption that the test results can be explained by the methods and the effects of the traits that represent L2 oral proficiency are negligible.

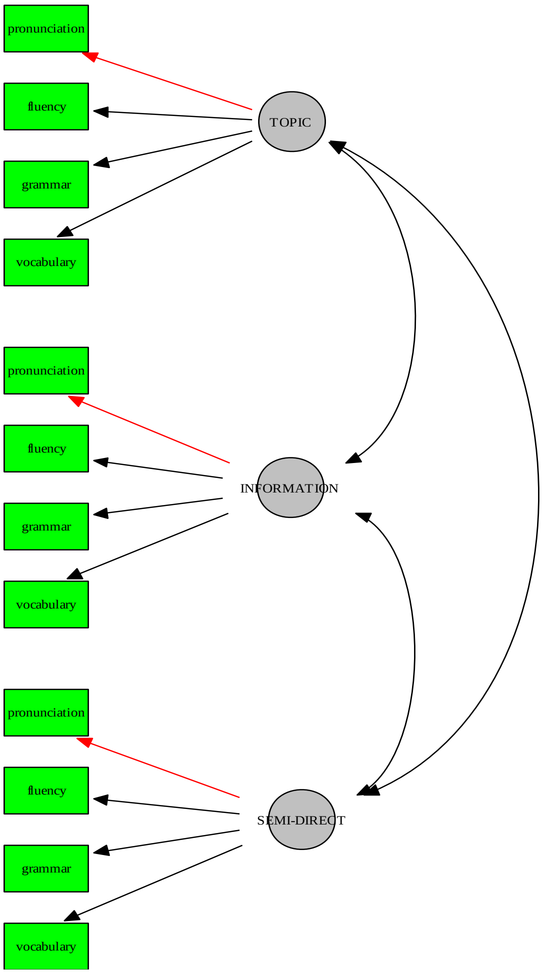

Figure 2. A no traits and freely correlated methods model (Model 2).

The second alternative model in Figure 3 (Model 3) assumes *perfectly correlated traits and freely correlated methods*. Technically, perfectly correlated traits can be achieved by setting the trait correlations equal to 1.0 in model specification. Theoretically, however, such a constraint suggests only one trait, denoted as *single trait* in the diagram, to be posited in the model as the traits are perfectly correlated. Finally, Figure 4 portrays the last alternative model of *freely correlated traits and perfectly correlated methods* (Model 4). In the model specification and calibration, the method correlations are set at 1.0. Such a constraint on method factors underlies the assumption that methods are not distinct from each other in assessing their corresponding traits.

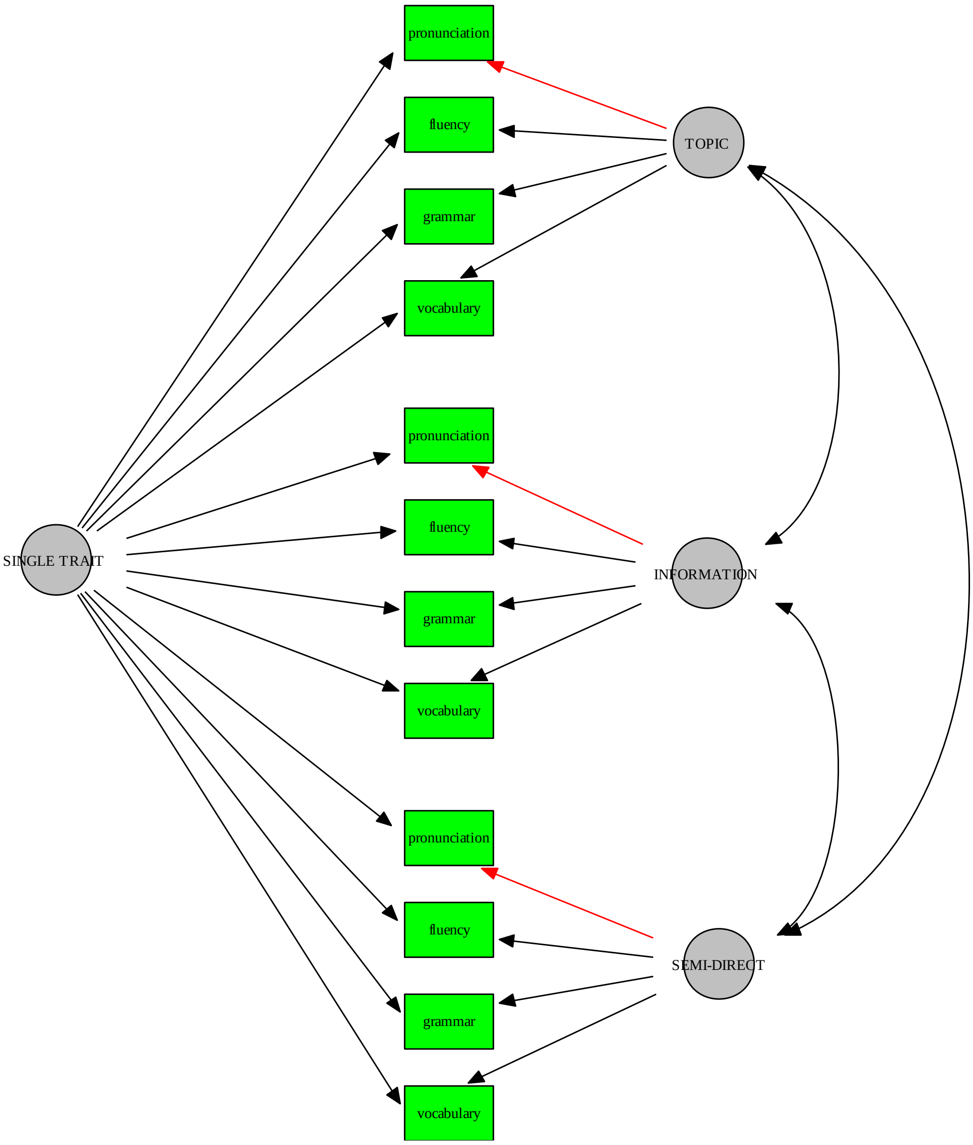

Figure 3. A perfectly correlated traits and freely correlated methods model (Model 3).

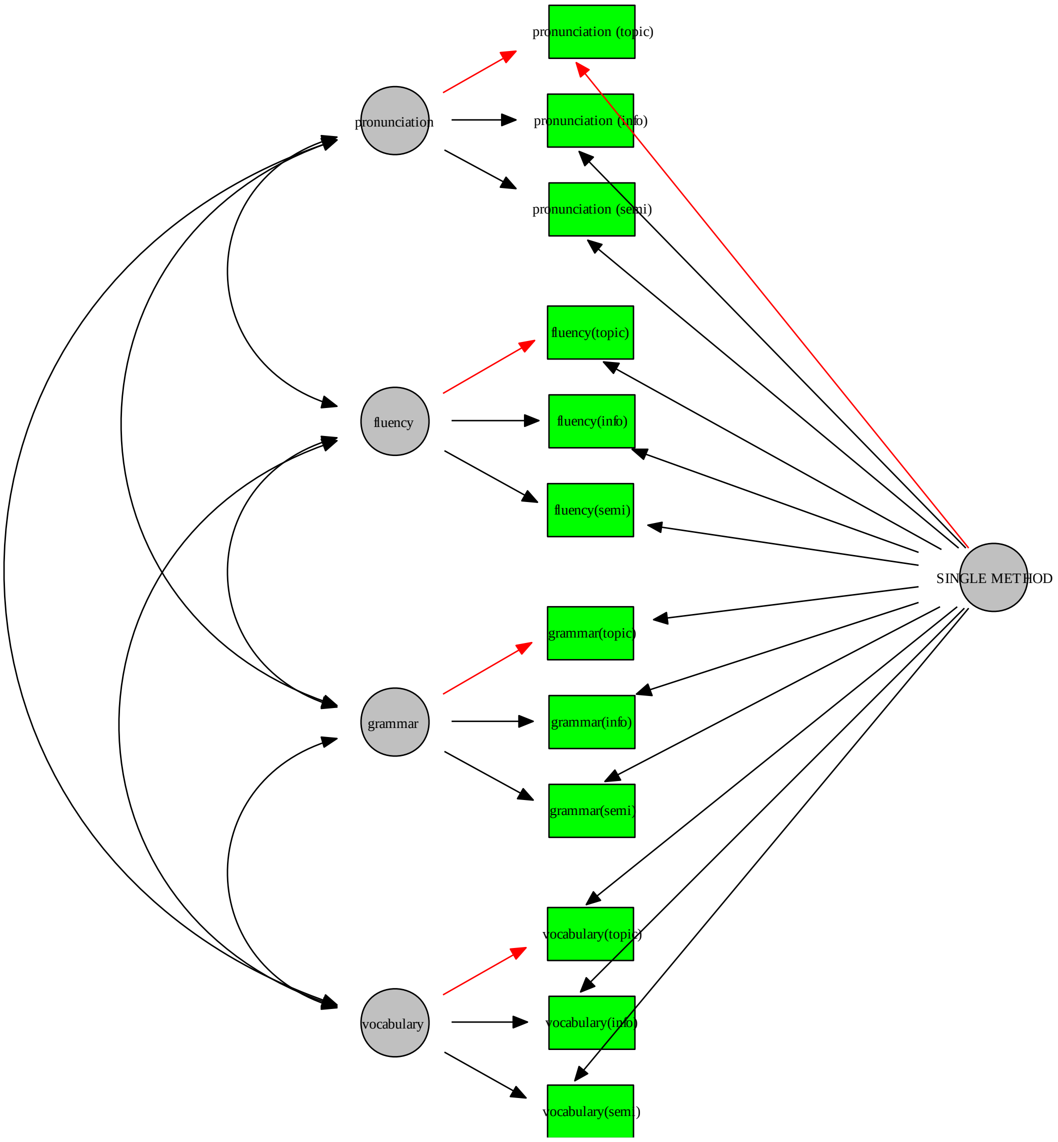

Figure 4. A freely correlated traits and perfectly correlated methods model (Model 4).

Table 4 presents a summary of the goodness-of-fit statistics related to the hypothesized and alternative models. As a couple of data variables were found non-normal, the entire data were analyzed using robust maximum likelihood (RML). Accordingly, as indicated in Table 4, fit indices estimated under RML are consulted and only the Robust Comparative Fit Index (RCFI) and the Bentler-Bonett Nonnormed Fit Index (NNFI) under Satorra-Bentler's scaled χ^2 statistic are reported. In discussing a good fit of a model to the data, the RCFI and NNFI indices of above 0.85 (or more conservatively 0.90) are considered desirable. As for the Root Mean Square Error of Approximation (RMSEA), Hu and Bentler (1991) suggested its value to be lower than 0.06 to argue for a good model-data fit.

TABLE 4
Summary of Goodness-of-Fit Indices for CTCM Models

Model	χ^2	*df*	*RCFI*	*NNFI*	*RMSEA*	*90%C.I.*
1. Freely correlated traits; freely correlated methods	42.094	33	.982	.969	0.02	.000, .061
2. No traits; freely correlated methods	94.047	51	.746	.672	0.07	.049, .096
3. Perfectly correlated traits; freely correlated methods	60.615	39	.873	.785	0.06	.027, .087
4. Freely correlated traits; perfectly correlated methods	70.413	36	.797	.628	0.07	.050, .105

Note. The $\chi^2$and *df* values are Satorra-Bentler's scaled, as robust maximum likelihood analysis was used.

To test convergent as well as discriminant validity of the traits and methods designated in the hypothesized model, the results of the χ^2 statistic and *RCFI* in Table 4 are consulted. The hypothesized model and each alternative model are compared pairwise with respects to their model-data fit. As Table 5 indicates, the difference of such a model fit between two competing models can be statistically determined for its significance using the χ^2difference ($\Delta\chi^2$) test (Byrne, 1994: 131). The discrepancy between the two models in comparison can also be evaluated non-statistically by checking the difference in practical fit, i.e., RCFI (ΔRCFI).

Evidence of convergent validity

Traditionally, evidence of convergent validity is confirmed when different measures of the same trait are highly correlated. In the CTCM CFA analysis, evidence of convergent validity can be determined by comparing a model in which traits are specified (Model 1 with freely correlated traits) with one in which they are not (Model 2 with no traits specified). This comparison serves to evaluate whether or not the addition of trait factors to the model helps explain a statistically significant amount of variance in the measurement. A significantly improved fit of Model 1 over Model 2 to the data, therefore, confirms the presence of convergent validity.

As Table 5 reports, the $\Delta\chi^2$ between Model 1 and Model 2 is 51.953 (*df* = 18), which is statistically significant at $p < 0.01$. Also, ΔRCFI is substantial at 0.236, thus confirming evidence of convergent validity between Model 1 and Model 2.

TABLE 5
Differential Goodness-of-Fit Indices for CTCM Model Comparisons

	Difference in		
Model comparison	χ^2	*df*	*RCFI*
Test of Convergent Validity			
Model 1 vs. Model 2 (traits)	51.953	18	0.236
Test of Discriminant Validity			
Model 1 vs. Model 3 (traits)	18.521	6	0.109
Model 1 vs. Model 4 (methods)	28.315	3	0.185

Note. The $\chi^2$and *df* values are Satorra-Bentler's scaled, as robust maximum likelihood analysis was used.

Evidence of discriminant validity

Evidence of discriminant validity can be evaluated based on both traits and methods. The discrimination between the trait factors supposes that they represent different constructs. In this matrix-level model comparison, a statistically significant model-data fit with Model 1 with freely correlated traits over Model 3 with perfectly correlated traits confirms the supposition that the four traits differ from each other and represent separate constructs for the identification of trait variance.

As shown in Table 5, Model 1 and Model 3 are compared to test for evidence of discriminant validity among the trait factors. When the $\Delta\chi^2$ value is significant favoring Model 1 with a non-negligible ΔRCFI, evidence of discriminant validity is confirmed. In Table 5, the significant $\Delta\chi^2$ value of 18.521 (df = 6, $p <$ 0.01) supports evidence of discriminant validity, which is additionally signified by the large size of ΔRCFI (= 0.109).

The same logic of investigation can be applied to the evaluation of discriminant validity with the methods. The significant results of the χ^2difference test favoring Model 1 with freely correlated methods over Model 4 with perfectly correlated methods indicate that the three methods specified in the measurement are substantially different from each other, separating method variance from that of trait and hence signifying the presence of discriminant validity.

In Table 5, the $\Delta\chi^2$ value of 28.319 between Model 1 and Model 4 is statistically significant (df = 3, $p <$ 0.01) and the size of ΔRCFI is substantial (= 0.185). Therefore, evidence of discriminant validity is strongly supported for the methods.

Testing for Construct Validity by Examining Parameters

A careful examination of individual parameter estimates can help draw a stronger argument for evidence of convergent and discriminant validity with the methods in the measurement of the traits. Although the hypothesized model in Figure 5 includes the factor loadings and correlations, to facilitate a more precise and organized discussion concerning the tenability of convergent and discriminant validity, the parameter estimates are reorganized and presented in two separate tables: Table 6 for trait and method loadings and Table 7 for trait and method correlations.

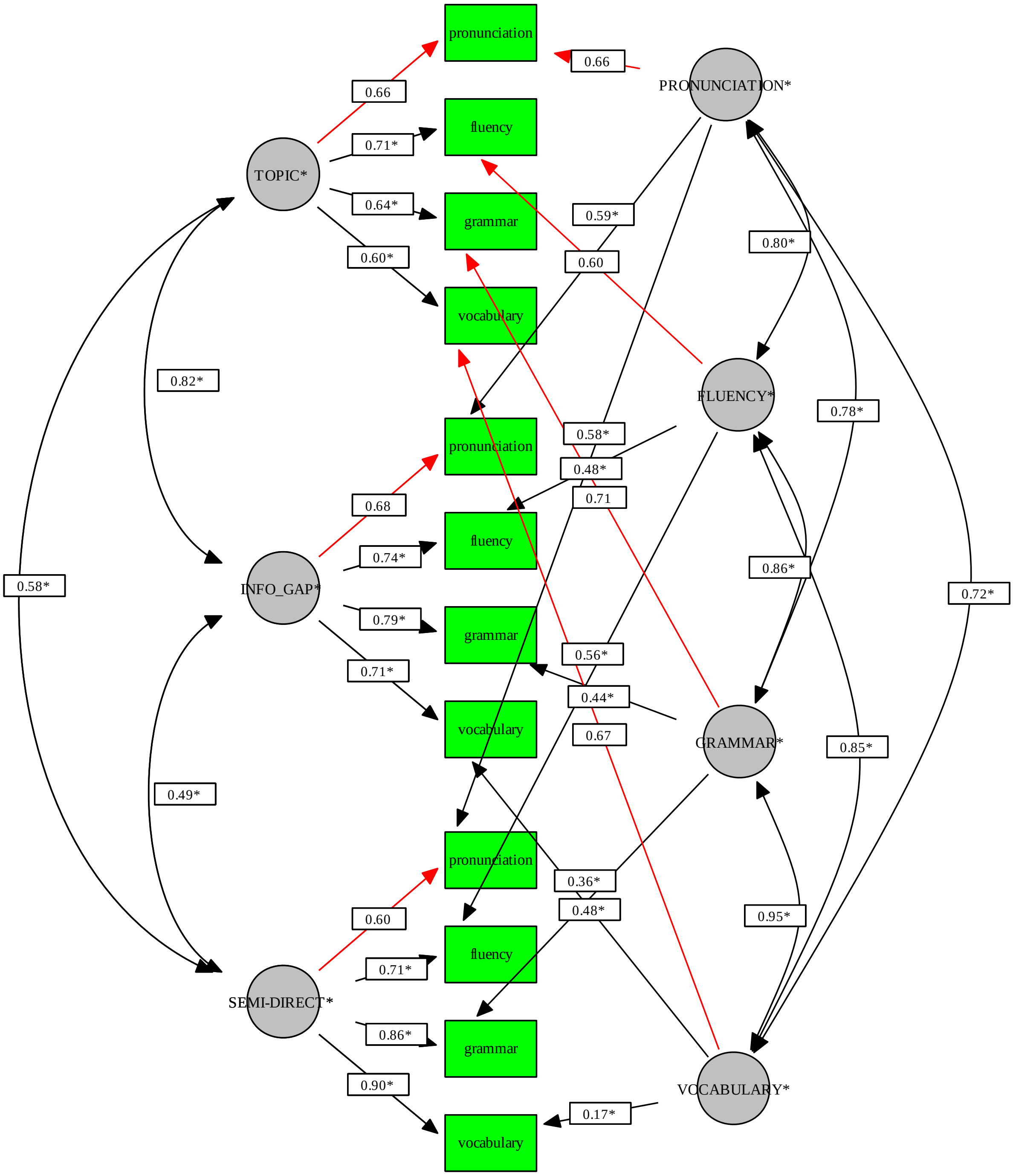

Figure 5. The hypothesized CTCM model with standardized estimates.

Evidence of convergent validity

To test evidence of convergent validity, the magnitude of trait loadings in Table 6 was examined. To argue for the validity, the trait factor loadings must be large, statistically significant, and larger than those of the method loadings.

As indicated in Table 6, all of the trait loadings are statistically significant with magnitudes ranging from 0.44 (Information gap of Grammar) to 0.71 (Topic discussion of Grammar) except for one loading of 0.17 (Semi-direct speaking of Vocabulary). However, in nine of 11 comparisons, the proportion of method variance exceeds that of trait variance, indicating a stronger method effect compared to that of the trait. Thus, evidence of convergent validity is tempered as the effect of traits is attenuated by that of the methods, especially related to information gap and semi-direct speaking.

At each trait level, the vocabulary trait using the semi-direct method was the most difficult to assess, while the same trait was assessed better using the topic discussion method. For the grammar trait as well, the topic discussion method worked best, while the information gap worked worst. For the traits of pronunciation and fluency, in the same way, the topic discussion method worked best in their measurement (average factor loading = 0.69). In short, as a method of trait measurement, the topic discussion contained the most trait variance, which is also indicated in its average factor loading of 0.66 compared to that of the method, 0.65.

TABLE 6
Trait and Method Loadings for the Hypothesized CTCM Model 1

	Methods			Traits				Comparison
	TD	*IG*	*SDS*	*P*	*F*	*G*	*V*	
Topic discussion (TD)								
Pronunciation (P)	0.66			0.66				Method/Trait
Fluency (F)	0.71				0.60			Method
Grammar (G)	0.64					0.71		Trait
Vocabulary (V)	0.60						0.67	Trait
Information gap (IG)								
Pronunciation (P)		0.68		0.59				Method
Fluency (F)		0.74			0.48			Method
Grammar (G)		0.79				0.44		Method
Vocabulary (V)		0.71					0.48	Method
Semi-direct speaking (SDS)								
Pronunciation (P)			0.60	0.58				Method
Fluency (F)			0.71		0.56			Method
Grammar (G)			0.86			0.63		Method
Vocabulary (V)			0.90				0.17[a]	Method

[a] statistically non-significant

Evidence of discriminant validity

Table 7 summarizes the information of the trait and method factor correlation matrices and is used to test evidence of discriminant validity. A strong argument can be made for the evidence if the factor correlations between traits are negligible.

TABLE 7
Trait and Method Correlations for the Hypothesized CTCM Model 1

	Methods			Traits			
	1	2	3	4	5	6	7
Method variable							
1. Topic discussion	1.00	--	--	--			
2. Information gap	0.83	1.00	--	--			
3. Semi-direct speaking	0.58	0.49	1.00	--			
Trait variables							
4. Pronunciation				1.00	--	--	--
5. Fluency				0.80	1.00	--	--
6. Grammar				0.78	0.86	1.00	--
7. Vocabulary				0.72	0.85	0.95	1.00

Note. All correlations are statistically significant at 5% level.

As the correlation matrix of trait variables in Table 7 illustrates, all six correlations are relatively high and statistically significant, ranging from 0.72 between the traits of pronunciation and vocabulary to 0.95 between the traits of grammar and vocabulary. Thus, discriminant validity is not demonstrated with the trait factors.

Additionally, evidence of discriminant validity with the method factors can be evaluated using the same type of information as provided in Table 7. Since employing multiple methods in a measurement presumes that the methods are distinct from each other, thus identifying different aspects of traits, negligible correlations between method factors signify evidence of discriminant validity. As reported in the table, the correlations between the semi-direct and each group oral method are marginal (0.58 and 0.49, respectively), thus ascertaining the presence of discriminant validity with the methods. To the contrary, the significant correlation of 0.83 between the two group oral methods of topic discussion and information gap suggests lack of discriminant validity.

Discussion

In appraising the comparability of the target tasks, this study first identified their unique characteristics using the framework (Table 1) by Pica et al. (1993), and then their construct equivalence was evaluated by demonstrating the presence of convergent and discriminant validity in their measurement of the traits. The construct equivalence was examined at the matrix level through model comparison and also at the individual parameter level that helps generate a stronger argument for evidence of convergent and discriminant validity.

In model comparison, the hypothesized CTCM model (Model 1) was compared against a nested series of more restrictive, alternative models (Models 2–4). Then, a series of χ2 difference tests were performed to determine their statistical significance. Evidence of convergent validity was tested by comparing Model 2 with Model 1, and the results of a significant Δχ2 value and a large size of ΔRCFI supported the presence of convergent validity with the traits specified in Model 1.

However, a more careful look at the individual parameter estimates of the hypothesized model revealed otherwise. Overall, the proportion of method variance was larger than that of trait variance, indicating that there was a stronger method effect compared to that of the trait. It was very much so with the traits under the two methods of information gap and semi-direct speaking, i.e., the trait effects were attenuated by the method effects, thus tempering evidence of convergent validity.

Evidence of discriminant validity for both traits and methods was tested and evidenced at the matrix level. The comparisons of Model 1 and Model 3 for traits and Model 1 and Model 4 for methods yielded significant Δχ2 and large ΔRCFI values, thus rendering support to discriminant validity. At the parameter level of factor correlations, however, the findings are mixed. While discriminant validity was not evidenced with the trait factors, it was with the methods, confirming the result from the matrix level analyses. Although the two group oral methods of topic discussion and information gap were highly correlated, neither of them was significantly correlated with the semi-direct method. The high correlation between the two group oral tasks may be the result of their format similarity that is yet to be confirmed.

The results of the current study suggest that the method effect in the model was relatively strong, and as a result, the trait was not fully reflected in the measurement process. Especially, the parameter level results indicated that the magnitude of the method effect was not consistent across different methods, with the topic discussion method containing the most trait variance. It helped produce more trait-related information compared to the other methods of information gap and semi-direct speaking. For discriminant validity, methods were found to be discriminant, while the traits were not.

Of particular note is the finding that although the correlation between the two group oral tasks was high, the parameter level analyses were able to demonstrate that each method assessed individual traits to differing degrees. The average factor loadings of topic discussion and information gap were 0.66 and 0.50 respectively, and the topic discussion method worked better for assessing all of the four traits compared to the information gap method.

The results of the study confirm the differential effects of the test tasks on examinee performance as reported by prior studies and suggest that the three methods of topic discussion, information gap, and semi-direct speaking not be considered comparable in assessing the L2 oral traits of pronunciation, fluency, grammar, and vocabulary.

Conclusions and Limitations

Considering the current trend in task-based assessment in L2 testing, more empirical research is required to promote understanding of the impact of interactive oral tasks on learner performance. Different tasks come with different strengths and weaknesses. Such aspects need to be continuously examined and addressed through empirical research findings.

In responding to such needs, the current study examined the comparability of three test tasks, topic discussion, information gap, and semi-direct speaking, mainly with respect to their construct equivalence using a CFA approach to CTCM design. The findings of this study contribute to the field of L2 assessment by adding new empirical evidence into the body of knowledge on test method effect.

Finally, several limitations need to be noted regarding the statistical techniques that this study employed for data analyses. First, the sample size entered into the CFA analysis is barely enough for a CFA study that involves the CTCM method. Also, the distributions of some score variables were non-normal and, consequently, the robust method had to be employed for the interpretation of the test statistics. As Hu and Bentler (1999) recommended, a sample size of over 250 observations may have helped produce more stable test statistics for a clearer interpretation of the research findings.

Acknowledgements

Part of this study is based on my doctoral dissertation (Park, 2008). The author would like to thank the Research Institute of Language Studies and Language Education (RILS) at Kanda University of International Studies for the funding in support of this research. He would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

The Author

Siwon Park is an associate professor in the Department of English at Kanda University of International Studies in Chiba, Japan. His research interests include second language acquisition, language assessment, psycholinguistics, and quantitative research methods.

Department of English
Kanda University of International Studies
Chiba, 261-0014, Japan
Tel: +81 432739763
Email: siwon@kanda.kuis.ac.jp

References

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning*, *31*(1), 67-85.

Bentler, P. M. (2003). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets* (Unpublished doctoral dissertation). King's College, University of London.

Bonk, W. J., & Ockey, G. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

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

Byrne, B. M. (1994). Burnout: Testing for the validity, replication, and invariance of causal structure across elementary, intermediate, and secondary teachers. *American Educational Research Journal, 31*(3), 645-673.

Byrne, B. M. (2006). *Structural equation modeling with EQS and EQS Windows: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Duff, P. A. (1993). Tasks and interlanguage performance: An SLA research perspective. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 57-95). Clevedon, Avon, UK: Multilingual Matters.

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal*, *30*(2), 156-167.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

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

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585–602.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning*, *33*(3), 315-332.

Hildon, J. (1991). The group oral exam: Advantages and limitations. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 189-197). London, UK: Modern English Publications and the British Council.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2). 177–204.

Luoma, S. (1997). *Comparability of a tape-mediated and face-to-face test of speaking: A triangulation study* (Unpublished licentiate thesis). Jyvaskyla, Finland: Jyvaskyla University Centre for Applied Language Studies.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests.* Frankfurt am Main, Germany: Peter Lang.

Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161–186.

O'Donnell, D., & Park, S. (2008). A FACETS analysis of the sub-traits of the KEPT oral rating scale. *The Journal of Kanda University of International Studies*, *20*, 385-404.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests* (Studies in language testing 13). Cambridge, UK: CUP.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295.

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

Pica, T., Kang, H-S., & Sauro, S. (2006). Information gap tasks: Their multiple roles and contributions to interaction research methodology. *SSLA*, *28*(2), 301-338.

Shohamy, E., Reeves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal*, *40*(3), 212-240.

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

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*(3), 275-302.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411-440.

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167-197.

Widaman, K. F. (1985). Hierarchically tested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*(1), 1-26.

Appendix

Oral Rating Scales

	Pronunciation Key points: ▪ **word level** ▪ **sentence level** (ability to 'blend' or link sound between words) ▪ **intonation**	**Fluency** Key points: ▪ **automatization** (ability to formulate utterances quickly and speak smoothly) ▪ **speaking speed** ▪ **hesitations and pausing**	**Grammar** Key points: ▪ **use of morphology** ▪ **complexity of grammar** (e.g., embedded clauses, parallel structures, connectors, etc.)	**Vocabulary** Key points: ▪ **range of vocabulary** ▪ **suitability of vocabulary use**
0.0 0.5	▪ Very heavy accent, that would lead to a breakdown in communication ▪ Only uses katakana-like phonology and rhythm ▪ Words not blended together	▪ Fragments of speech ▪ Halting, often incomprehensible ▪ Communication nearly impossible	▪ Shows severely limited grammar knowledge	▪ Knows few words or phrases, and uses them in isolation
1.0 1.5	▪ Uses somewhat Katakana-like pronunciation ▪ Does not blend words ▪ Likely to have comprehension difficulties with interlocutors	▪ Slows strained, unnatural speech ▪ Frequent unnatural groping for words ▪ Long unnatural pauses ▪ Communication difficult	▪ Shows some very limited grammar knowledge ▪ Does not have enough grammar to express an opinion clearly; makes frequent errors ▪ Makes no attempt at complex grammar	▪ Uses lexical forms not adequate for task ▪ Cannot express opinion properly with the limited words used
2.0 2.5	▪ Have not mastered some difficult sounds of English, but would be mostly understandable to interlocutors ▪ Make some attempt to blend words	▪ Speech is hesitant ▪ Unnatural groping for words and unfilled spaces may persist, but does not completely impede communication ▪ May overuse fillers, or demonstrate other unnatural usages	▪ Overly reliant on a small range of simple grammar ▪ Has enough morpho-syntax to express meaning ▪ Complex grammar is attempted, but may be inaccurate	▪ Generally has enough vocabulary for expressing some opinion, but does not demonstrate any advanced/special knowledge of vocabulary
3.0 3.5	▪ Pronunciation is good, but has still not mastered all the sounds of English ▪ Accent does not interfere with comprehension ▪ Can blend words consistently	▪ Occasional misuse of fillers and groping ▪ Frequent repair may still be evident, but is not overly distracting	▪ Shows some ability to use complex grammar ▪ May make errors, but they are only in late-acquired grammar ▪ Errors do not impede meaning	▪ Shows evidence of some advanced/special vocabulary, but they are used inaccurately and unnaturally

4.0	▪ Shows excellent pronunciation and intonation ▪ Has practically mastered the sound system of English	▪ Conversation should proceed smoothly, with little impediment ▪ Use fillers, markers, & lexical chunks effectively	▪ Uses both simple and complex syntactic structure effectively ▪ Major errors are rarely noticeable	▪ Shows evidence of a wide range of lexical forms, and uses them accurately

Note:

- If a student shows she is consistently fulfilling the criterion being tested, she receives the lower score in the box; if she *sometimes* achieves the expected level, but sometimes slips to lower criteria, she is given the higher score in the box of the lower criteria (band).
- If a student did not speak enough for you to reliably assign a score for a category, see if you can get them to speak more. If they don't oblige, assign them **U** as a score for that category.