Journal of Pan-Pacific Association of Applied Linguistics, 21(1), 45-63.

The Effects of Test Method on L2 Reading and Listening Performance

Siwon Park
Kanda University of International Studies

Park, S. (2017). The effect of test method on L2 reading and listening performance. *Journal of Pan-Pacific Association of Applied Linguistics, 21*(1), 45-63.

This paper examines how different test methods may tap different aspects of second language knowledge. It employs multiple-choice (MC) and constructed response (CR) items which yield distinct or convergent information in the computer delivered testing of English in its presentation of this factor. In order to examine the effects of test method, a CFA approach to a multitrait-multimethod design was adopted to examine the convergent as well as discriminant validity of the two skill area traits and the two test methods. After comparing the hypothesized model with a nested series of more restrictive models using χ^2 difference tests (Byrne, 2006; Widaman, 1985), based on the information from the path coefficients, and also using principal factor analyses, it was found that: 1) the skills are factorially separable, yet, highly correlated, 2) MC and CR methods are separable, and very weakly correlated, 3) there is thus a method effect, and 4) the MC method may not be suitable in sampling the unique ability characteristics of reading and listening skills.

Keywords: test method, method effect, MTMM, L2 reading and listening

1 Introduction

Bennett and Ward (1993) address many of the assumed limitations of multiple-choice (MC) assessments, as they "encourage the teaching and learning of isolated facts and rote procedures at the expense of conceptual understanding and the development of problem-solving skills" (ix). They further argue that, on the contrary, several potential limitations are associated with the use of constructed response items such as fewer questions being asked, lower content coverage, lack of standardization, and lack of score comparability due to more subjective criteria in evaluation. However, it is still not clear whether constructed response items provide unique information on the focal trait that is not captured by multiple-choice items.

1.1 Multiple-choice vs. constructed response

The literature concerning test method indicates that the disjunction between choice and construction is often exaggerated and becomes unclear when we

underscore the processes of examinee responses rather than the test formats. As such, researchers generally posit different test methods on a continuum from multiple-choice on one end to construction/presentation on the other (Bennett, 1993).

Method effect is often examined in light of two different, yet somewhat related facets in measurement: validity and practicality. Work by Bennet (1993), Wainer and Thissen (1993), and others indicates that multiple-choice questions provide essentially the same information as constructed response questions and that multiple-choice questions are superior because more of them can be given in the same amount of time and at less cost. It is also suggested that multiple-choice questions can be administered and scored quickly, still maintaining high reliability and objectivity in scoring. In addition, multiple-choice questions enable the testers to sample broader knowledge domains. In some cases, however, the equivalence of construct coverage in the two item types occurs simply because the constructed response items are the open-ended versions of the multiple-choice items. Thus, those items may be simply measuring the same low-level knowledge as that in the original multiple-choice items.

The criticism that multiple-choice questions are restricted to fairly low level and non-complex types of information has been the motivation for those who promote the use of the constructed response format. They argue that multiple-choice questions are not able to assess knowledge involving higher order cognitive skills. This leads to a possible threat to the validity of the inferences from the test score given that the construct on a multiple-choice test may be under-represented. Another possibly negative impact of the frequent use of multiple-choice questions in educational measurement may be that in a large-scale, standardized testing setting, teachers and learners may target the domains and skills only required to answer multiple-choice questions, often regarded as teaching to the test format. This criticism has been Frederikson's (1984, 1990) position for some time. Also, Ward, Frederikson, and Carlson (1980) and Frederikson, Ward, Case, Carlson, and Samph (1981) found that multiple-choice and constructed response questions measured different constructs. Birenbaum and Tatsuoka (1987) noted considerable differences between the two formats, with more favorable results for the open-ended format. In their review of SLA research findings, Norris and Ortega (2003) also noted that the selective response types did not correctly reflect the complex nature of SLA research and the interpretation of its findings. Unlike these research findings favoring the use of construct response format, however, researchers such as Stevens and Clauser (n.d.) and Ackerman and Smith (1988) argue that both formats should be included, especially for language tests, as their studies reveal a significant effect of method.

Mixed findings in prior studies on test method effect may be in part due to a heavy focus on item formats rather than on the cognitive skills

required in responding to the test items. Researchers such as Bennett (1993), Snow (1993), and Messick (1993) argue that method effect in assessment should not be seen as a one-dimensional continuum from responses that are more selective to those more constructed. There are other aspects as to method effect that testing and SLA researchers need to be concerned about since it directly concerns test validity. It is important to consider method effect from multiple perspectives including test formats, cognitive skills required to respond to an item, affective factors with the test conditions, and stakes involved, in addition to the knowledge and skills to be measured. For instance, there may be items that require skills more cognitively invariant; items not easily affected by the response method. There may be cases in which a multiple-choice item requires a more cognitively demanding process to reach the correct answer and other instances in which a constructed response item can be answered by activating a simple recall process.

1.2 Method effect in language testing

In language testing, a line of studies has been concerned with the impact of method effect on examinee performance, while another has examined how test method interacts with other test variables such as examinee attributes and/or text characteristics and affects test performance.

Bachman and Palmer (1981) examined the impact of three testing methods of oral interview, reading translation, and self-rating on the measurement of reading and speaking skills. They employed Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach as well as confirmatory factor analysis (CFA) and found strong method factors affecting the traits of interest. In a study using three oral testing methods of imitation, completion, and interview, Henning (1983) also observed that the methods might not all be measuring the same aspect of oral proficiency. Bae and Bachman (1998) conducted a study of English language tests employing a latent trait approach and examined whether reading and listening skills were factorially separable. They indicated the two skills were separable, yet highly correlated.

In a study that concerns the interplay between test method facets and text features and/or learner characteristics, Shohamy (1984) investigated the effects of multiple-choice and open-ended questions in L1 and L2 on students' reading comprehension of the same L2 texts. She found that method, language, and text all significantly influenced students' scores and especially, the three variables influenced students of low proficiency the most. Likewise, Kobayashi (2002) examined the effects of text organization and response format on L2 learners' performance in reading comprehension tests and found test format as well as test organization had a significant impact on their performance. She also noticed that the impact was not consistent across the levels of learner proficiency – only the learners' performance of high

proficiency was affected by different response formats. Similarly, Lee (2011) examined the extent to which response format, text difficulty, and proficiency levels interact with each other making differential impacts on EFL students' performance on a reading test. He also discovered a complex interplay among the three test variables affecting test performance.

On the listening side, Cheng (2004) examined to what extent the form and type of multiple-choice and open-ended response influence the test takers' listening performance. He noticed that the different types of responses had a significant effect on test takers' performance. They performed the best on the test of multiple-choice format and the worst on the open-ended.

Finally, In'nami and Koizumi (2009) conducted a meta-analysis of the effects of multiple-choice and open-ended formats on L2 reading and L2 listening test performance. Their analyses informed that multiple-choice formats were easier than open-ended in L2 listening; yet, no format effects were notable in L2 reading except when the studies involved L2 learners of high proficiency and/or employed design features such as between-subjects, random assignment, or stem-equivalent items.

As reviewed thus far, a number of studies have examined method effect in L2 reading and listening performance. However, it is rare to find ones that 1) employ the MTMM CFA approach (rather than the Campbell and Fiske's (1959) MTMM approach[1]), 2) concern both L2 reading and listening and cross-examine test performance in relation to method effect, and 3) examine L2 performance on the computer delivered tests. Therefore, this study, using the CFA approach to MTMM analysis, examines the extent to which multiple-choice and constructed response items are distinct or convergent in the computer-delivered test format across the two skill areas of L2 reading and listening.

[1] Earlier MTMM studies often employed the Campbell and Fiske's (1959) approach based on bivariate correlation analysis. However, as the interpretations of the trait and method variables and the strength of their relationships are subjective, such studies often suffered from inconclusive findings: the subjective nature of the judgments led the researchers to draw unprecise validity arguments as to the relationships between traits and methods. The CFA approach to MTMM analysis helps overcome this limitation as it enables the researchers to examine the relationships among the supposed latent traits and methods simultaneously accounting for the variance due to each effect of trait, method, and error. The MTMM CFA approach provides statistical means for such examinations in terms of model comparisons using χ^2 difference tests as well as factor loadings and correlations.

2 Method

2.1 Participants

The data for this study come from 145 Japanese learners of English at a university in Japan, 92% of them majoring in English and the rest in International Communication. Approximately 50% of the 145 participants were enrolled in their first year, 30% in the second year, and the rest in their 3rd and 4th year of study. Hence, their age ranges approximately from 18 to 22. Among them, 70% were female students and the rest male students. The students volunteered to participate in this study by taking the tests and answering a questionnaire about their test-taking experiences and perceptions about the test formats and items.

2.2 Test instruments and data collection

The reading and listening items in the tests were developed based on a range of item functions identified from the Common European Framework of Reference–Japan (CEFR-J) Reading and Listening scales (Tono, 2013). Both easy and difficult items were included in the test so that a range of differing abilities could be sampled. The tests were piloted and revised, and the final versions included 20 selected response items and eight constructed response items for each skill.

All items were presented in the same format. The multiple-choice items had a format similar to that of TOEFL reading test. Reading passages and listening texts were developed to represent the CEFR-J levels of A 1.3 through B2.2, and each of them accompanied two to five test items depending on their length and text difficulty. Each item consisted of a stem and four choices, one being the correct answer and the rest being the distractors. The skills required to answer the questions include getting the main idea, specific information, and inferencing, and vocabulary only for reading.

The constructed response items had a format of one passage (or one script in case of a listening test) and one question. The examinees were required to write a lengthy answer to each question. Once the responses to the constructed responses were submitted, two raters of highly qualified EFL teachers scored each response based on a set of scoring guidelines. Only when a large discrepancy occurred between the two scores from the raters on the same response, a third rater evaluated the response again and resolved the discrepancy by taking the average of the two closer scores. The rater agreement between the first two raters for all of the response samples resulted in $r = .837$.

The tests were delivered on the computer. Originally, 178 students took the tests for this study for over two years; however, those incomplete responses were deleted *listwise*, and only complete response sets from 145 students were entered in the analyses, resulting in a 19% data loss.

2.3 Design

As noted earlier, the primary purpose of the current study is to examine the effects of multiple-choice and constructed response formats on L2 reading and listening test performance. In order to achieve the purpose, we adopted a CFA approach to MTMM analysis and examined the extent to which the trait and method structure of the test data demonstrated the convergent as well as the discriminant validity and method effect. Convergent validity refers to the extent to which different methods agree in measuring the same trait, while discriminant validity concerns the extent to which different methods diverge in measuring different traits (Byrne, 2006). As presented in Figure 1, the hypothesized model is a 2 x 2 MTMM model. By comparing the hypothesized MTMM model against a nested series of more restrictive models (Byrne, 2006; Widaman, 1985), the present study explores the factorial relationship between the traits across the methods. That is, in this CFA approach to the MTMM analysis, the hypothesized model is compared to all alternative models with respect to which model best statistically fits the data. Alternative models are constrained by trait, method, or both.

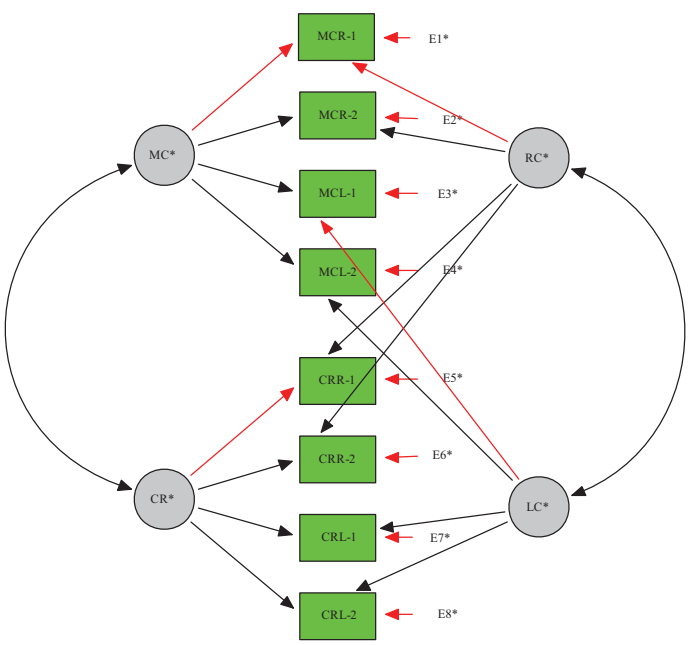

Figure 1. Hypothesized MTMM CFA model (Correlated traits/correlated methods)
Note: LC: listening comprehension RC: reading comprehension
MC: multiple-choice CR: constructed response
MCL: multiple-choice listening tests CRL: constructed response listening test
MCR: multiple-choice reading tests CRR: constructed response reading test

On the right side of the model in Figure 1 are the two trait factors, listening and reading; on the left side are the two method factors, multiple-choice and constructed response methods. In addition, the remaining four alternative models are presented in Figures 2 through 5.

The first of the more restrictive models, presented in Figure 2, is a model with a unitary trait (or perfectly correlated traits) with no method factor included in the model. Thus, the model is restricted to a single trait and no methods. Figure 3 represents a model with no traits, but with freely correlated methods. The model in Figure 4 again has a restricted unitary trait, with the methods freely correlated, as opposed to a single method or no method. Figure 5 presents the opposite model of that in Figure 4, with two freely correlated traits and perfectly correlated methods. Finally, the model in the five models including the hypothesized 2 x 2 MTMM model were each identified and their χ^2 goodness-of-fit indices were calculated for the validity discussion in the following section. These χ^2 *difference tests* provide evidence regarding the convergent and divergent validity, and an examination of the factor loadings is to reveal method effects if present.

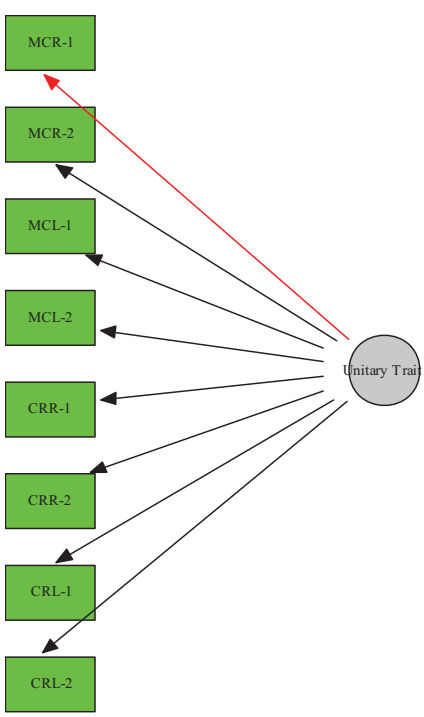

Figure 2. A perfectly correlated traits and no methods model

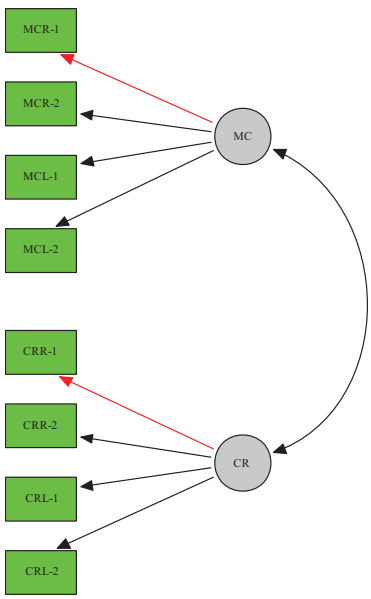

Figure 3. A no traits and freely correlated methods model

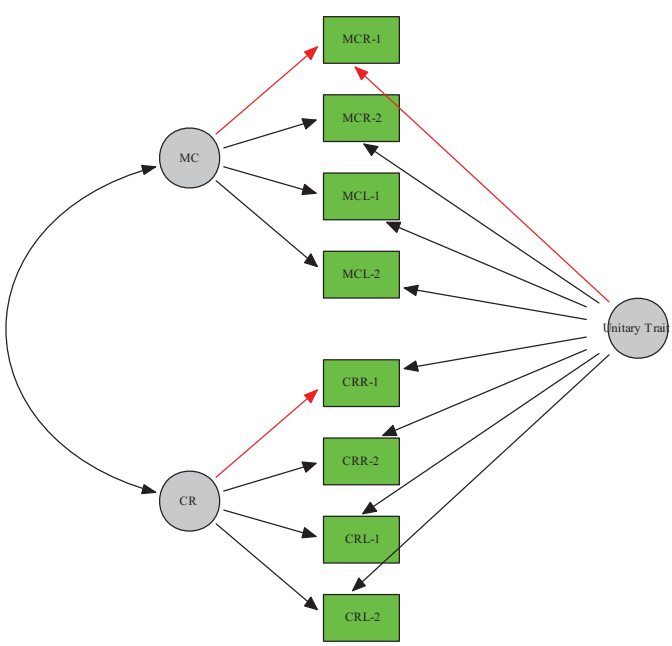

Figure 4. A perfectly correlated traits and freely correlated methods model

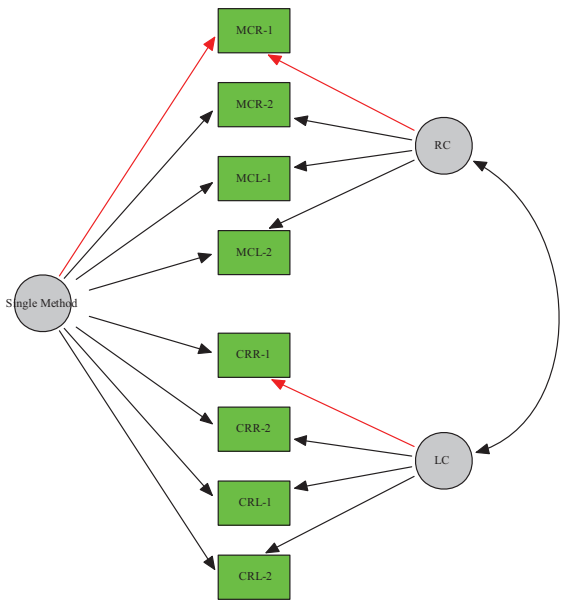

Figure 5. A freely correlated traits and perfectly correlated methods model

3 Results and Discussion

All analyses were done using IBM SPSS 21.0 (IBM Corp, 2012) and EQS 6.1 for Windows (Bentler, 2004). Using the covariance structure (see Appendix for the covariance and correlation matrices) based on the observed scores, specified models were estimated for their model fit indices and factor loadings. Table 1 presents the descriptive statistics and the information about the reliability of each section of the tests and the distribution of the data.

Table 1. Descriptive Statistics and Reliability and Data Distribution Indices

Tests	*k*	*M*	median	*min.*	*max.*	*SD*	alpha	*Skew.*	*Kurt.*
RC MC 1	10	6.30	6	0	10	2.55	.78	-0.76	-0.38
RC MC 2	10	6.68	7	1	10	2.40	.76	-0.15	-0.74
LC MC 1	10	6.78	8	0	10	2.83	.83	-0.18	-1.04
LC MC 2	10	6.28	7	0	10	2.74	.81	-0.32	-0.88
RC CR 1	4	7.58	7	0	18	4.82	.81	-0.71	0.41
RC CR 2	4	7.47	8	0	18	4.61	.80	-0.40	0.40
LC CR 1	4	10.87	11	0	18	3.66	.83	0.40	-0.60
LC CR 2	4	10.23	10	0	19	4.87	.85	-0.71	-0.20

Note. *N*=145; *RC*: Reading comprehension; *LC*: Listening comprehension
MC: Multiple-choice; *CR*: Constructed response

Most of the alpha values exhibit relatively high reliability except those from the two multiple-choice reading comprehension (RC) sections with comparatively lower values of .78 and .76. Also, note that the alphas from RCs are generally lower than those from LCs. With respect to the univariate distributions of the data based on skewness and kurtosis values, most of the test sections demonstrate negligibly low skewed and peaked distributions (within the range of ±1), indicating that the distribution of the data is close to normal and hence satisfies the assumption of the parametric statistical analysis.

3.1 MTMM model comparisons

In testing CFA models, it is customary to first test the most restrictive model. That is, if there is no reason to reject a single factor explanation, then there is generally no reason to explore more complex models (Kline, 2005). Subsequently, the remaining models are examined in the order from the least restrictive postulated model through the more restrictive models. Consequently, we first tested the model in Figure 2. Then, the models in Figures 3, 4, and 5 were compared to the hypothesized model in Figure 1.

Table 2 presents the summary of goodness-of-fit indices for the five models from Figures 1 – 5. Generally, χ^2 values that are not significant indicate model fit, while significant values indicate that the model does not fit the data. However, it is known that CFA is sensitive to sample size and tends to produce larger values of χ^2 as the sample size increases. This may lead to somewhat spurious interpretations of model fit. Therefore, an adjustment of the χ^2 values is made by dividing the observed χ^2 values by the degrees of freedom (χ^2/df) (Kline, 2005). A suggested ratio criterion of 3 or below indicates model fit. Also, conventionally, the CFI, NFI, and NNFI fit indices should all be above .90 in order to indicate fit.

In the analysis presented in Table 2, the χ^2 value for Model 1 is too high, demonstrating a statistical significance, the χ^2/df ratio is above 3, and the CFI, NFI, and NNFI are all below .90. These findings indicate that the model does not fit the data. Therefore, this model was not explored further.

The fit indices for the other four remaining models including the MTMM model are also presented in Table 2 for model comparisons using the χ^2 *difference tests* which were used in Table 3 to perform model comparisons. As shown in Table 2, only the χ^2/df value for Model 2 is below 3. The ratios of Models 3, 4, and 5 are significant, being higher than the criterion value of 3.

Table 2. Summary of Goodness-of-fit Indices for the Models

Model	χ^2	*df*	χ^2/df	*NFI*	*NNFI*	*CFI*
1. Freely correlated traits; Freely correlated methods	28.62*	10	2.86	.97	.94	.979
2. Perfectly correlated traits; No methods	382.15*	20	18.25	.59	.44	.600
3. No traits; Freely correlated methods	169.02*	19	8.90	.82	.76	.834
4. Perfectly correlated traits; Freely correlated methods	43.58*	11	3.96	.95	.91	.964
5. Freely correlated traits; Perfectly correlated methods	73.32*	11	6.67	.92	.83	.931

* sig. $p<.05$

Table 3. Differential Goodness-of-fit Indices for MTMM Model Comparisons

Model comparisons	χ^2	*Difference in df*	*CFI*
Test of Convergent Validity			
Model 1 vs. Model 3 (traits)	140.40*	9	.145
Test of Discriminant Validity			
Model 1 vs. Model 4 (traits)	14.96*	1	.015
Model 1 vs. Model 5 (methods)	44.70*	1	.048

* sig. $p<.05$

According to Byrne (2006), convergent validity in the MTMM design can be explored by estimating the extent to which independent measures of the same trait are correlated and improve the model description. This can be tested by comparing a model that includes traits and methods with another that includes only methods. This is the comparison indicated in Table 3 between Model 2 and Model 3. The large improvements in the χ^2 value ($\Delta\chi^2$ = 140.40) and the CFI index (ΔCFI = .145) support the presence of convergent validity of the two skills in the MTMM model. For a demonstration of discriminant validity, two comparisons were made: one for traits and the other for methods. For the discriminant validity of the traits, Model 2 was compared with Model 4 which represents a unitary trait (or perfectly correlated traits). The significant difference of the χ^2 values ($\Delta\chi^2$ = 14.96) between the two models indicates that it is conceivable for the two traits (i.e., language skills) to be factorially separable. However, the improvement of model fit between the two models indicated by the value of CFI is rather marginal (ΔCFI = .015) and does not fully support the argument for the discriminant validity between the two models. Therefore, the aspect of discriminant validity is fully observed due to the minimal improvement through specifying two separate traits rather than perfectly correlated ones.

The same procedure was taken to demonstrate the discriminant validity with the test methods by comparing Model 2 against Model 5. The comparison resulted in a significant improvement in the χ^2 value ($\Delta\chi^2 = 44.70$) and a sizable increase in CFI (ΔCFI = .048) suggesting the presence of discriminant validity. Discriminant validity, in this case, is indicated by the improvement due to specifying two methods rather than perfectly correlated methods.

3.2 Interpretation of the factor loadings

The factor structure for both the MC and CR items was examined for similarities and differences in the way the items loaded on different factors. Factor loadings may be similar or dissimilar between those for MC and others for CR. Table 4 and Figure 6 present information with the factor loadings results from the CFA analysis for the hypothesized MTMM model.

Table 4. Factor Loadings for Each Path Analyzed (Standardized Estimates)

Variables		*Methods*		*Traits*			
		MC	*CR*	*RC*	*LC*	*E*	R^2
MC	RC1	.48		.65		.59	.66
	RC2	.22		.80		.57	.68
	LC1	.72			.65	.25	.94
	LC2	.61			.66	.43	.81
CR	RC1		.89	.27		.36	.87
	RC2		.90	.21		.38	.85
	LC1		.48		.74	.48	.77
	LC2		.60		.71	.36	.87

Note. *RC*: Reading comprehension; *LC*: Listening comprehension
MC: Multiple-choice; *CR*: Constructed response

If the corresponding loadings are similar, they concur in their measurement of the same underlying ability. First, the dissimilarity of the factor loadings obtained between MC and CR method measuring each skill indicates that method effect is present. For instance, the MC method appears to do a better job measuring listening comprehension with the listening loadings of .72 and .61 than reading comprehension. Likewise, the constructed response method works better for reading comprehension, with the reading loadings of .89 and .90.

Overall, there appears to be convergent validity present as the factor loading for trait measurements are high ranging from .65 to .80, except the two low factor loadings, .27 and .21 from the RC trait to the two CR reading measurements. The attenuation of the trait effects is caused by the strong CR methods as they temper the effects of the trait onto the two CR reading measurements.

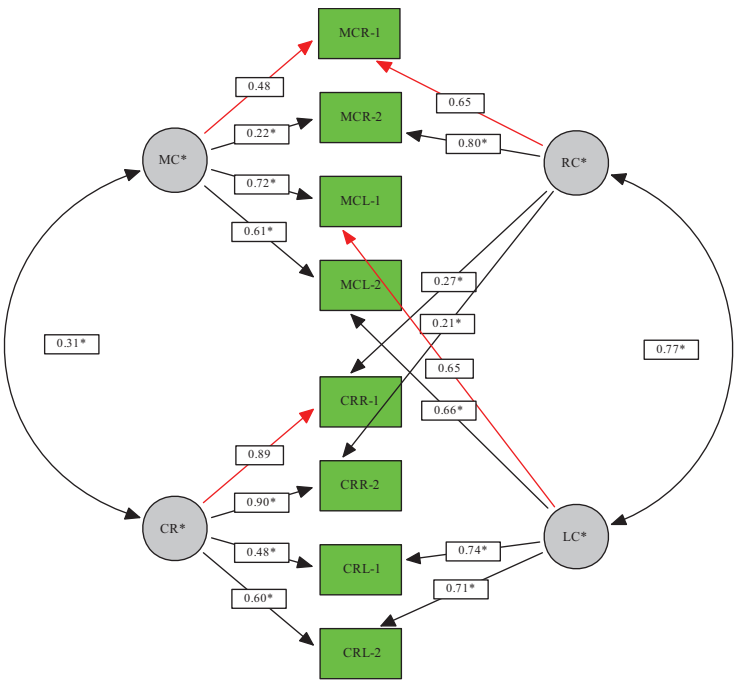

Figure 6. The hypothesized MTMM model with path coefficients

Note. LC: listening comprehension RC: reading comprehension
MC: multiple-choice CR: constructed response
MCL: multiple-choice listening tests CRL: constructed response listening test
MCR: multiple-choice reading tests CRR: constructed response reading test

Table 5. Trait and Method Correlations for the Hypothesized MTMM Model[a]

	Traits		*Methods*	
Measures	RC	LC	MC	CR
Reading Comprehension (RC)	1.00			
Listening Comprehension (LC)	**0.77**	1.00		
Multiple-choice (MC)			1.00	
Constructed Response (CR)			**0.31**	1.00

[a] Standardized estimates

Note. All correlations are statistically significant

As for the trait and method correlations shown in Figure 6 and Table 5, the relatively low correlation between the two methods demonstrates method effect – i.e., the evidence of discriminant validity. On the contrary, the high correlation between the traits and the coefficients between the traits and observed variables indicates the presence of convergent validity. These

findings are in line with the earlier observations of convergent as well as discriminant validity demonstrated by a χ^2 difference test between the models.

3.3 Principal factor analysis

To better understand what possibly had caused such an interacting effect between the two skills across two methods revealed by the earlier MTMM analyses, principal factor analyses (PFA) were performed with the data. PFA helps to discover joint variations among observed variables in response to latent variables; hence, it enables us to examine the factor structures of the observed responses and check whether or not test methods are responsible for such joint variations.

First, a PFA was performed with the CR data and the structure was rotated so that we could examine the data better by amplifying the factorial contribution of each item to the factors produced. The extraction method produced three distinctive factors as shown in Table 6. Essentially, the LC items mostly loaded on the first factor while the RC items on the second factor, indicating that the constructed response method successfully differentiated examinees' performance between the two skills. The third factor was rather spuriously related either to the LC trait or to the RC trait. The component plot in Figure 7 demonstrates that the reading and listening CR items are placed apart from each other in rotated space.

Table 6. Rotated Component Matrix for the Data from the CR Sections

Rotated component matrix			
		Factors	
Items		1	2
RC items	1		.685
	2	.442	.665
	3	.423	.588
	4		.771
	5		.809
	6		.790
	7		.794
	8	.566	.454
LC items	9	.683	
	10	.654	
	11	.736	
	12	.842	
	13	.662	
	14	.851	
	15	.671	.439
	16	.775	

Note. Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization

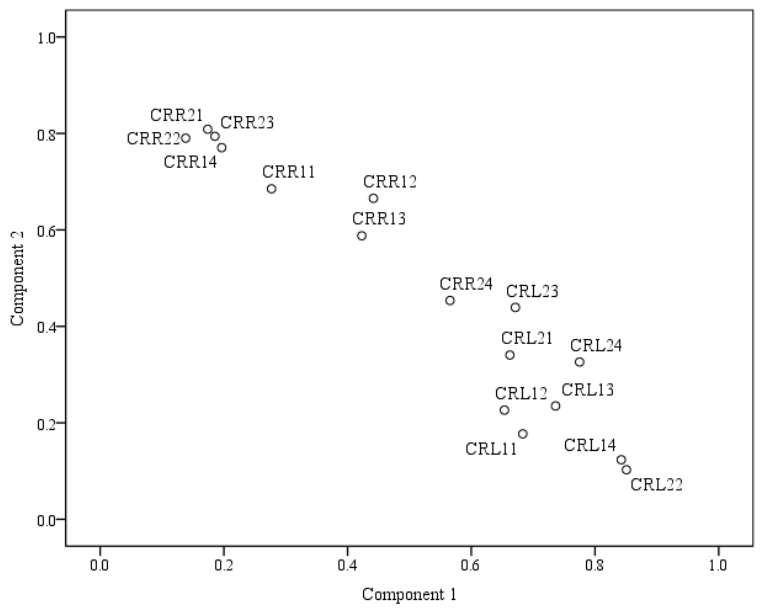

Figure 7. Component plot for CR items in rotated space
Note. CRR – reading items; CRL – listening items

Another PFA was performed this time with the MC data. Four MC sections basically resulted in 12 factors with an eigenvalue higher than 1.00. The 12 factors may have resulted as a function of a relatively large number of MC items; yet, 12 factors from 40 items appear too many. This diffusive characteristic of MC items indicates that the MC method does not differentiate the two possibly distinct skills to any important degree. That is, the MC method, as it requires a lower order of cognitive skill applications, does not appropriately function to sample out the distinctive characteristics of each skill. Figure 8 confirms the diffusive characteristic of MC items scattered on the rotated space without skills separation.

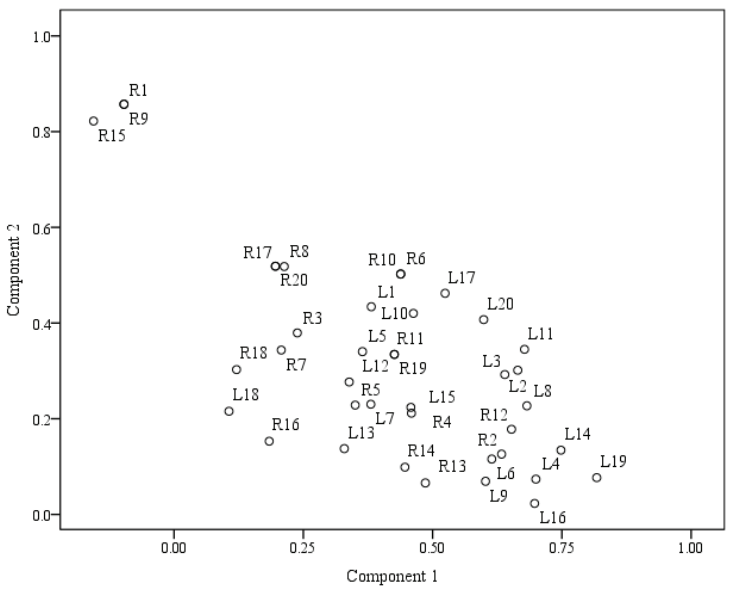

Figure 8. Component lot for MC items in rotated space
Note. CRR – reading items; CRL – listening items

4 Conclusions

Using a series of χ^2 difference tests among the proposed models, we have demonstrated the convergent validity for the traits and the convergent as well as discriminant validity for the methods. As the correlation between the two skills demonstrated in Figure 6, listening and reading are separable, but highly correlated (r =.77). This is similar to the findings in Bae and Bachman (1998). The two methods, MC and CR are clearly separable as they are weakly correlated (r =.31). This result also confirms the existence of a method effect. However, the two methods differentially predict scores on listening and reading. The multiple-choice section is a better predictor for the listening comprehension section while the constructed response format works better for the reading comprehension.

A follow-up factor analysis was performed to see the predictability of each method in measuring the two skills. CR scores resulted in two factors. The first factor was mostly predicted by the observations from the RC tests and the second factor was mostly from the LC tests. This appears to indicate that CR is a method that can appropriately predict different skills when the multi-skills are present in a set of test items, as in RC and LC in our tests. This finding is not surprising considering the low correlation between the two methods. However, as the component plot for the MC items in rotated space demonstrates, the MC method appears less sensitive to the skills that items

are to sample and hence was not able to differentiate them. Therefore, the MC method may not be suitable in sampling the unique ability characteristics of reading and listening skills.

As the findings of the study indicate, test methods directly influence the results of L2 reading and listening skills assessments. That is, the manipulation of test formats in such assessments could bear a significant influence on the validity arguments as to the use of the tests. Therefore, the development of L2 reading and listening tests requires careful consideration into the response format which would enable a more accurate assessment of L2 comprehension.

References

Ackerman, T. A., & Smith, P. L (1988). A comparison of the information provided by essay, multiple-choice, and free-response tests. *Applied Psychological Measurement, 12*(2), 117-128.

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

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the English/English two-way immersion program. *Language Testing, 15*(3), 380-414.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment*. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Bentler, P. M. (2004). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4)*,* 385-395.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-104.

Cheng, H. F. (2004). A Comparison of multiple-choice and open ended formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544–555.

Frederikson, N. (1984). The real test bias. *American Psychologist, 39*(3), 193-202.

Frederikson, N, (1990). Introduction. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. ix-xvii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Frederikson, N., Ward, W. C., Case, S. M., Carlson, S. B., & Samph, T. (1981). *Development of methods for selection and evaluation in undergraduate medical education.* Princeton, NJ: Educational Testing Service.

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

IBM Corp. Released 2012. *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing*, *19*(2), 193-220.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. NY: The Guilford Press.

Lee, J-W. (2011). A Comparison of Constructed Response Formats as Measures of EFL Reading Comprehension. *English Teaching*, *66*(2), 149-167.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment* (pp. 61-74). New Jersey: Lawrence Erlbaum Associates, Inc.,Publishers.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717-761). Oxford: Blackwell.

Shohamy, E. (1984). Does the Testing Method Make a Difference? The Case of Reading Comprehension. *Language Testing*, *1*(2), 147-70.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and*

portfolio assessment (pp. 45-60). New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Stevens, J. J., & Clauser, P. (n.d.). *Multitrait-multimethod comparisons of selected and constructed response assessments of language ability.* Retrieved from http://www.uoregon.edu/~stevensj/papers/mtmm.pdf. November, 2008.

Tono, Y. (2013). *CAN-DO list creation and the use of the CEFR-J guidebook.* Tokyo: Taishukan.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and multiple-choice versions of a test. *Journal of Educational Measurement*, *17*(1), 11-29.

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

Siwon Park
Kanda University of International Studies
1-4-1 Wakaba, Mihama-ku, Chiba-shi Chiba, 261-0014, Japan
Phone: 81-43-273-1913
Email: siwon@kanda.kuis.ac.jp

Received: April 30, 2017
Revised: July 5, 2017
Accepted: July 10, 2017