Research Note: Construct validity of the Pearson Test of English Academic: A multitrait-multimethod approach[1]

Hye K. Pae, Ph.D.
University of Cincinnati, U.S.A.
September 2012

Linguistic abilities are multifaceted and cannot easily be measured directly. Consequently, multiple diagnostic methods are employed to assess multiple traits of language skills. This indirect means of assessment raises issues of adequacy, appropriateness, and utility of the measures in testing an individual's ability or skill. With the contribution of Campbell and Fiske (1959) to the field of language testing, a multitrait-multimethod (MTMM) approach has been used by many researchers not only to validate inferences made based on test scores, but also to build theories. The MTMM approach is largely used to examine how different traits (multitraits) of language abilities and methods (multimethods) of testing materials influence the student's performance. Campbell and Fiske's (1959) original conceptualization was characterized by hypotheses that correlations among scores measuring a unitary ability (monotrait correlations) would be higher than those among scores using a single test method (monomethod correlations), and monotrait-monomethod correlations would be higher than those among measures of different traits using different methods (heterotrait-heteromethod correlations). This stipulation has received criticism. The main criticism points to the ambiguity of the magnitude of adequate correlation coefficients and the reliance on correlations that do not allow for quantification of the amount of the specific variance in the data to make inferences about underlying dimensions, such as trait and method factors (Bachman, 2004; Marsh, 1989; Widaman, 1985). Other abilities or test methods may affect the test-taker's performance. Another criticism is related to failure to separate method variance from random error in the MTMM correlation matrix (Brown, 2006; Schmitt & Stults, 1986). Although a comparison of the magnitude or strength of the correlations across traits and methods provides valuable information, a challenge lies in the absence of criteria to determine the magnitude of correlations necessary to claim significant differences (Bachman, 2004).

A confirmatory factor analysis (CFA) can be used to overcome the limitation of a simple comparison of MTMM correlations, as CFA models are a powerful and direct means to test the relative contributions of traits and methods to test-takers' performance and to explain underlying relationships (Bachman, 2004). A CFA model with multiple-trait factors and multiple-method factors specifies factor loadings of different measures on their associated traits and method factors as well as zero loadings on all other factors.

1. The Purpose of the Study

The aim of this study was to evaluate the validity of linguistic constructs and assessment method effects utilizing an MTMM matrix. Of interest was an assessment of convergent validity, discriminate validity, and the effect of method variance in the field test of the Pearson Test of English Academic (PTE Academic). Three discrete constructs and one integrated-skill construct were conceptualized

[1] This research note is partly drawn from Pae (2012).

as traits, including listening, reading, speaking, and integrated skills, and each construct had three indicators. Each construct was assessed using three different methods: prescribed multiple-choice question format, constructed question format, and summarized question format. The criterion that differentiated the constructed test method from the summarized one was made based on the flexibility the test-taker enjoyed in relation to the prompt provided. Specifically, the constructed method required the test-taker to generate an answer beyond the available parent text; therefore, the argument and meaning of the generated portion can be variable according to examinees' responses. On the other hand, the summarized method asked the test-taker to shorten the given text within the restricted word limit.

A CFA model was conceptualized based on two hypotheses as follows:

> Hypothesis 1: Four separable, correlated, language performance constructs (i.e., traits: listening, reading, speaking, and integrated skills) will show substantial convergent validity and discrimant validity in relation to the methods used.
>
> Hypothesis 2: The method variance will be insignificant, if it is present.

2. Method

2.1 Participants and Measure

This is a secondary data analysis. Five hundred eighty-five examinees' test scores were selected from the score database of the second field test of PTE Academic. The participants were adult English language learners (ELLs), and their mean age was 25 years, ranging from 17 to 59 years of age. Females accounted for 54.2% and males 45.8%. According to their self-report, 53% of the participants had studied English for more than 10 years, and 57% had lived in English-speaking countries. PTE Academic assesses ELLs' overall English skills as a second language (L2) or a foreign language (FL)[2], covering real-life English used in English-medium academic settings. The instrument measured a range of ELLs' English skills using a mixture of various item types and formats.

2.2 Variable Building and Analysis

Since each section of PTE Academic assessed different skills using different task types (i.e., independent language proficiency, such as listening, reading, speaking, and integrated skills), one aim was to examine how these task types affected test scores and score validities of the assessment. Four traits (listening, reading, speaking, and integrated skills) and three methods (prescribed, constructed, and summarized) were concurrently analyzed.

First, an MTMM correlation matrix was obtained to examine convergent validity, discriminant validity, and construct validity. Next, a CFA correlated traits and correlated methods (CTCM) analysis was performed. The CTCM model consisted of four correlated language constructs and three correlated method factors.

[2] Since it is not known whether the participants learned English as a second language (L2) or foreign language (FL), L2 and FL are used interchangeably in this paper for the sake of convenience and consistency in the literature.

3. Results

3.1 Convergent Validity, Discriminant Validity, and Construct Validity: An MTMM Correlation Matrix

An MTMM correlation matrix of the variables under consideration was examined. Convergent validity coefficients indicate correlations between the scores of the same trait using heteromethods. Discriminant validity coefficients, which are typically smaller than those of convergent validity, indicate correlations between scores of different traits using the same method. Table 1 displays a multitrait-multimethod correlation matrix. The correlations among the different traits (i.e., listening, reading, speaking, and integrated skills) are nested within each assessment method. The monotrait-monomethod correlation coefficients (i.e., reliability; the internal consistency of subscores on the instrument) are shown in parentheses. The boldface represents estimates of monotrait-heteromethod correlation (i.e., construct validity), indicating that different methods of theoretically congruent constructs are strongly interrelated. The underlined coefficients show heterotrait-monomethod correlations, while the regular typefaces indicate heterotrait-heteromethod correlations. The methods of theoretically different traits using different methods provide evidence of discriminant validity in comparison to the monotrait blocks. According to Campbell and Fiske (1959), the monotrait-monomethod correlation (reliability) and the monotrait-heteromethod correlation coefficients (validity) should be higher than those of heterotrait-monomethod and heterotrait-heteromethod correlations. The lower heterotrait-heteromethod coefficients (i.e., correlations between subscores that share neither trait nor method) than those of validity diagonals represent convergent validity which is significantly different from zero and sufficiently large enough to call for further examinations of validity (range of rs: .23 - .71). This satisfies the first requirement of convergent validity according to Campbell and Fiske's (1959) criteria. The other requirement is that the validity diagonal values should be higher than the values of heterotrait-heteromethod and heterotrait-monomethod coefficients. This condition, which is discriminant validity, is modestly met.

Since direct comparisons of the diagonal values to the heterotraits blocks demonstrated an inconsistent pattern, the means of the coefficients by the trait-method blocks were computed. The average of reliability (monotrait-monomethod) coefficients was .57, while that of validity (monotrait-heteromethod) coefficients was .42. Although the reliability coefficients were not uniformly higher than those of validity, it showed convergence of the independent methods. Although the correlations within each monotrait-heteromethod triangle block (i.e., validity) were not systematically larger than the heterotrait-monomethod correlations, the average (.42) of validity diagonal coefficients was also higher than that (.40) of the heterotrait-monomethod correlations, indicating convergent validity. There was partial validation, according to Campbell and Fiske's (1959) requirement, that the validity diagonals exceed the heterotrait-heteromethod control values. The average of heterotrait-momomethod correlation coefficients was higher than that of the heterotrait-heteromethod correlations, indicating discriminant validity (.40 and .37, repectively).

The presence of method effects was examined through the off-diagonal values of the monomethod blocks. Some method variance was observed, although not great, especially for the prescribed-response measure, compared to the constructed-response method. Specifically, the elevation of reading2 from reading1 indicated the presence of the method variance (i.e., difference between r = .18 and r = .32). This finding was consistent with the reliability coefficients. The mean of the prescribed reliability coefficients was lower than that of the constructed counterpart (.55 and .60, respectively). The inspection of the differences among the assessment methods used was useful for predicting common method bias because the correlations partially determined the covariance among the different methods. In short, Hypothesis 1 was supported.

Table 1. Multitrait-Multimethod Correlation Matrix (n=585)

	1	2	3	4	5	6	7	8	9	10	11	12
Prescribed												
1. Listening1	(.37)											
2. Reading1	.18	(.02)										
3. Speaking1	.39	.23	(.95)									
4. Integrated1	.47	.30	.76	(.86)								
Constructed												
5. Listening2	**.35**	.21	.55	.57	(.42)							
6. Reading2	.32	**.28**	.40	.44	.45	(.56)						
7. Speaking2	.43	.23	**.67**	.62	.43	.40	(.71)					
8. Integrated2	.40	.30	.62	**.71**	.56	.50	.56	(.69)				
Integrated												
9. Listening3	**.39**	.28	.57	.63	**.40**	.44	.49	.54	(.67)			
10. Reading3	.31	**.23**	.30	.36	.28	**.34**	.34	.42	.40	(.49)		
11. Speaking3	.30	.23	**.54**	.50	.32	.31	**.31**	.42	.41	.20	(.69)	
12. Integrated3	.21	.13	.57	**.45**	.27	.21	.21	**.32**	.36	.15	.50	(.40)

Note: All correlations are significant at the .01 level.

Parentheses: monotrait-monomethod correlations; reliability coefficients
Underlined: heterotrait-monomethod correlations
Boldface: monotrait-heteromethod correlations
Regular typeface: heterotrait-heteromethod correlations

3.2 CFA Approaches to the MTMM Matrix

A CFA specification of correlated traits correlated methods (CTCM) was performed. The three methods (i.e., prescribed response, constructed response, and summarized response) were used to gauge four traits (i.e., listening, reading, speaking, and integrated abilities) of L2 academic English skills (see Figure 1). The factor pattern for the CTCM model was expected to provide an account for observed relationships among the series of test scores. Figure 1 displays the CTCM model in which the two sets of traits and the method variables are correlated with one another within the category, but the different traits are uncorrelated with the different methods.

The initial CTCM model did not converge. A start value for initial parameter values was specified as 1.0, and admissibility check (AD = OFF) was set in order to obtain parameter estimates. For model modification, the latent variable variances were set to 1.0 and factor correlations between traits and methods were set to 1.0 in order to avoid the nonpositive definite Phi matrix. Twelve error variances were also set to zero to prevent the Heywood case. The correlations among the trait factors and the method factors were freely estimated, but the correlations between the trait and method factors were set to zero. With the model modification through admissibility check and a priori set variances, the model was a modestly appropriate fit [χ2 (33, N = 585) = 91.63, p = .000, χ2/df = 2.78, CFI = .99, GFI = .97, RMSEA = .056] (see Table 2).

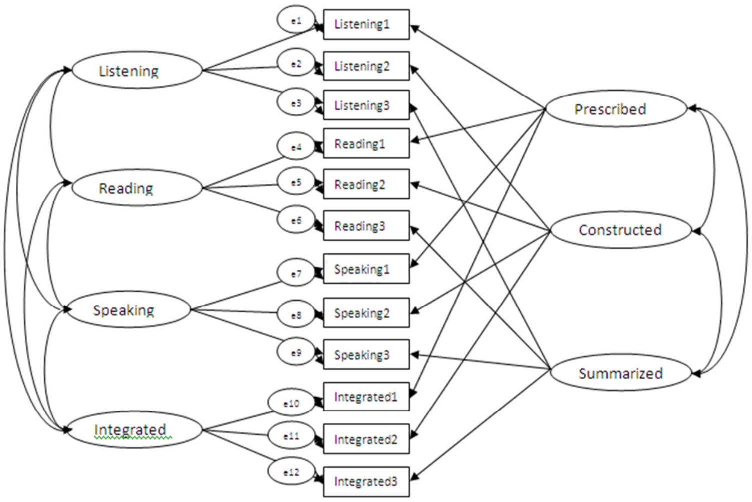

Figure 1. A Model for the Multitrait-Multimethod Correlated Traits and Correlated Methods

Table 2. MTMM Standardized Factor Loadings of Four Traits using Three Methods (n=585)

	Traits				Methods			
	Listening	Reading	Speaking	Integrated	Prescribed	Constructed	Summarized	Error Variance
Listening1	.51				.11			.73
Listening2	.60					.25		.57
Listening3	.68						.11	.53
Reading1		.40			.06			.84
Reading2		.63				.28		.52
Reading3		.51					.15	.71
Speaking1			.88		-.03			.23
Speaking2			.76			.04		.42
Speaking3			.63				-.16	.57
Integrated1				.90	.12			.17
Integrated2				.77		.37		.28
Integrated3				.64			-.72	.07

Note: The latent variable variances were set to 1.0 and factor correlations between traits and methods are set to 1.0 for model identification.

The variance of the listening trait factor using the multiple-choice question format accounted for 26% (.512 = 26%), while the method variance played a minimal role (.112 = 1%), with the error variance of .73. As these three elements were independent sources of variance, the three values equaled 100%. Given that the squared trait communality of the average factor loadings for the integrated skills accounted for 59% (mean factor loading = .77), the trait factor seemed to be the primary source of the variance in the integrated skills. The reading trait factor explained only 26% (mean factor loading = .51) of the variance associated with the measured variable. The larger trait factor loadings than those of the methods indicated convergent validity. The assessment of reading1 had the highest error variance when using the multiple-choice item type (error variance = .84), indicating that reading was the most difficult trait to assess using the forced-choice assessment method. The second most difficult trait to measure using the multiple-choice item type was the listening trait (error variance = .73). For listening, the item type which required constructed responses worked best (standardized factor loading = .25, error = .57). The constructed-response item type worked best for reading assessment as well (standardized factor loading = .28, error = .52). For integrated skills, however, the summarized responses showed the highest magnitude of factor loading but the direction was negative (standardized factor loading = -.72, error = .07). In short, the reading trait showed the highest mean error (mean error variance = .69), while the integrated skills showed the lowest mean error (mean error variance = .17). Thus, the reading skills were the most difficult trait to assess using any of the three methods, especially with the forced-choice method (factor loading = .06; error variance = .84). These findings were consistent with those found in the analysis of the MTMM correlation matrix. As indicated by the small method-factor loadings, Hypothesis 2 was supported.

4. Discussion

This study investigated the validity of linguistic constructs in ELLs' performance on PTE Academic, using an MTMM approach. Since educational inferences and decisions made on the basis of test scores of the assessment instrument have significant consequences for test-takers, an examination of the effects of traits and methods on test performance is important. Test effects on ELLs cannot be ignored because they may have different cultural and cognitive sets from those of native English speakers and because the norming samples used in the standardization procedure for high-stakes tests typically under-represent ELLs due in part to the continuous influx of ELLs into the English-speaking countries.

Although there is a theoretical basis for distinguishing language skills, each skill is not clearly distinct and independent. The construct-related approach for test validity is crucial because it focuses on the role of theory or conceptual framework in test construction and on the need to formulate hypotheses that can be examined as part of the validation process. Construct-related investigations are, in general, comprehensive because they involve content relevance and representativeness as well as psychometric evidence.

Overall, the results confirmed that ELLs' English performance was primarily influenced by the trait factors and that the language achievement traits were only partly influenced by the question format utilized. According to Campbell and Fiske (1959), it is possible that many multitrait-multimethod matrices show no perfectly convergent validation in real data. The findings of this study show that the assessment methods are adequate, mainly for measuring the given traits, and the question formats do measure the postulated traits. Since the traits of language proficiency are multicomponential and do not show a functional unity, it is possible that the test-taker's response tendency involved is specific to the construct-irrelevant attributes of each test, such as item layouts and font sizes.

The results provided evidence that some portion of the variance was related to the three question-formats, suggesting that the question type might have assessed different constructs, especially for the prescribed-response format. Specifically, the result demonstrated that the reading1 indicator had the highest error variance with the forced-choice question type, followed by the listening1 multiple-choice method. One explanation relates to unique qualities associated with the multiple-choice format which requires application, deduction, and evaluation of concepts. In a multiple-choice test, the examinee can work backward from multiple-choice answer options to figure out or guess a correct answer. This problem-solving strategy is specific to multiple-choice questions, and is not applicable to the other forms of questions, such as constructed-response or essay format. This distinctive characteristic of the multiple-choice format might explain the high error variance of the given indicator. Another possible explanation has to do with a wide range of item difficulty that the prescribed forced-choice question can cover. Since the indicator was an aggregated score of the multiple-choice responses, the source of the high error variance was undetectable in this study. The findings of this study suggest questionable validity of the multiple-choice question format that is designed to measure ELLs' reading and listening proficiency, despite the economic advantages of the question type. Irrespective of the source of the method effect, the multiple-choice question type assessing reading and listening skills calls for special attention. On the other hand, the constructed-response format seems to be a comparatively efficient means to assess ELLs' English skills.

Since validation process is an ongoing effort (Linn & Miller, 2005), the results of this study contribute to the field with respect to the provision of empirical evidence for ELLs' linguistic traits and assessment method effects presented in PTE Academic. Moreover, the different sources of test input (e.g., visual and aural prompt, scripted aural passage, etc.) were not considered in this study. Importantly, different question types employ different textual contents with different readability levels. A future study that investigates the impact of test input on test performance, controlling for the readability level of passages used in the instrument, will provide new insights into method effects.

To conclude, since any attempt to identify ELLs' strengths and weaknesses should stem from theories of language acquisition, learning processes, and usage, any approach to diagnosis and high-stakes decisions in relation to ELLs' academic English performance must take account of research evidence in language performance. This study adds empirical evidence to the extant body of knowledge with respect to the relationship between language traits and test methods. This kind of study needs to be continued, because most language tests are intended to cater to diverse learners. Therefore, a number of different text modes, question item types, and test methods should be used, not only in order to address test-takers' multifaceted skills and affective features, but also to assess the different attributes of language skills. In spite of the method effect found in the multiple-choice question format measuring reading and listening, PTE Academic appears to properly measure ELLs' English academic skills with respect to the critical qualities of language tests, such as construct validity, reliability, authenticity, consequences, interactiveness, and practicality, which eventually contribute to test usefulness (Chappelle, Jamieson, & Hegelheimer, 2003).

1 References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). Language Teaching, 35, 79-113.

Bachman, L. F. (2004). Statistical analyses for language assessment. New York, NY: Cambridge University Press.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York, NY: Guilford.

Campbell, D. T., & Fiske, D. W. (1959). Covergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Chappelle, C. A., Jamieson, J. M., & Hegelheimer, V. (2003). Validation of a web-based ESL test. Language Testing, 20, 409-439.

Linn, R. L., & Miller, M. D. (2005) (9th ed.). Measurement and assessment in teaching. Upper Saddle River, NJ: Merrill Prentice Hall.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. Applied Psychological Measurement. 13, 335-361.

Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson Test of English Academic. Educational Research and Evaluation, 18, 211-229.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

Schumacker, R. E., & Lomax, R. G. (2004) (2nd ed.). A beginner's guide to structural equation modeling. New York, NY: Taylor & Francis Group.

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