TELL, Vol.2, No.8, 2008

The Construct Validity of a Language Proficiency Test: a Multitrait Multimethod Approach

Abbas Ali Rezaee
Assistant professor, Tehran University

Mohammad Salehi
Assistant professor, Sharif University of Technology

Abstract

The present study intends to investigate the construct validity of a test which is used for admission purposes. It was conducted with 3,398 participants. These testees took an English language proficiency test as a partial requirement for entering a PhD program in different fields of education at the University of Tehran. This test has three sections consisting of grammar, vocabulary and reading comprehension. To determine the construct validity of the test, the design of multitrait-multimethod (MTMM) was investigated. According to Cronbach and Meehl (1959), in this design and approach to validation more than one trait and more than one method should be used. In the present study, the two traits used were grammar and vocabulary. The two methods employed were multiple choice and contextualization. The results revealed that the test possessed both convergent and discriminant validities. In other words, the same traits tested through different methods had more correlation than different traits tested through the same methods. To collect more evidence for the validity of the test and determine the effect of the methods used, the researchers also conducted two monotrait multimethod analyses: one was grammar being tested through more than one method and the

other was vocabulary being tested in more than one way. The effects of methods were also determined.

Key words: construct validity; multitrait-multi-method; convergent validation; discriminant validation; language proficiency test

I. Introduction

Test validation is an important enterprise, especially when the test is a high-stakes one. The University of Tehran administers a proficiency test to PhD candidates on a yearly basis. The test is a high stakes one; almost 9000 candidates take the test. High-stakes assessment is any assessment whose outcome has life-changing implications for the test taker. Admission tests for universities or other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations (Roever, 2001).

Acceptance into the PhD program hinges on the results of this test. High stakes tests need to be validated. If the validity of a test like this one is not known, it might have undesirable consequences for the society at large (Messick, 1988).

So far to the knowledge of the researchers no in-depth study has been conducted regarding the validity of the test. There has been only one MA thesis directed at the validity of the test (Zand Karimi, 2005). But, we believe the study is insufficient as it does not deal with the test in its totality. Only the reading comprehension section was exposed to investigation, while the other sections, too, need an in-depth analysis.

The current study adopts multitrait-multimethod (MTMM) approach to investigate the construct validity of the aforementioned test.

2. Review of the Related Literature

2.1 Approaches to Construct Validation

There have been several approaches to test validation. Alderson, Clapham, and Wall (1995) came up with the following approaches to construct validation. The first approach that they touch on is the correspondence with theory. In other words, the test results are supposed to confirm the theory. According to Alderson et al the theory itself is not called into question. The second approach they bring to the readers' attention is internal correlations. If a test battery is composed of some sub-parts, like a proficiency measure, then the correlations of these sub-parts should be low, so that evidence can be collected on the distinctness of these parts. The authors rightfully assert that the correlation of any sub-part with itself is necessarily one or perfect. Now, to assure that the test has construct validity, the subparts should be correlated with the total test. Still, another problem may arise; the correlation of any sub-part with the total test with the sub-part included on the total test may inflate the correlational index. To solve that problem, the authors suggest excluding that particular sub-part from the total test and then running the correlation. Still, another approach they elaborate on is factor analysis. MTMM approach which is the concern of the present study is also elaborated. Finally, the last approach is that of taking into account test bias and actually assessing the role of background knowledge, gender, race, etc.

As MTMM designs as a tool for investigating construct validity are the prime concern in the present study, they are required o be fully elaborated on.

Perhaps the pioneers for MTMM designs can be Campelle and Fiske (1959). Palmer and Groot (1981) maintain that the design was applied to language testing by Stevenson (1981). There will be an overview of the concept followed by theoretical underpinnings to be further followed by research studies.

Test scores may be the function of the trait and the method used to test it. For example, a trait may be tested differently by different methods like multiple choice completion and simple completion. If two individuals with the same overall grammatical knowledge perform differently under the two test conditions using two different methods, then the difference can be attributed to the influence which using different methods has exerted. Essential to the MTMM designs are the notions of *convergent* and *divergent* validity.

As to the convergent validity, it can be said that if a trait is to be tested by two methods, because the trait is the same in each method, the correlation is expected to be high. So, if a group of testees take a grammar test in the form of multiple choice and simple completion, the correlation is supposed to be high because in each case grammar is being tested and any difference can be attributed to the effect of the method.

On the other hand, divergent or discriminant validity is logically related to the convergence of scores. The difference between convergence and divergence can be illustrated with an example. Vocabulary and grammar are supposed to tap different constructs. To the extent that these two produce a low correlation speak to the discriminant validity of the tests.

Palmer and Groot (1981) rightfully remind us that a high correlation between two apparently distinct traits may indicate that the two may be related deep down. For example, reading and writing are supposed to be distinct traits and a low correlation is expected. But a relatively high correlation goes to show the two skills tap similar skills like vocabulary knowledge and world knowledge. As Palmer and Groot maintain the MTMM designs can be shown in a matrix. To illustrate the point, the example pointed out above can be shown by a matrix as in Table 1.

Table 1: An Example of an MTMM Design

Methods / Traits	Multiple Choice	Fill-in-the-blank
Grammar	Test #1 :Multiple Choice test of grammar	Test # 2: Fill-in-the blank test of grammar
Vocabulary	Test # 3: Multiple choice test of vocabulary	Test # 4: Fill-in-the-blank test of vocabulary

(Taken from Palmer & Groot, 1981, p.7)

As it can be observed, the two traits (grammar and vocabulary) and the two methods (multiple choice and fill-in-the blanks) are shown in the matrix. Correlational analysis can provide evidence for the convergent and discriminant validity of the tests. High correlations between test #1 and test # 2 will provide evidence for the convergent validity of the grammar tests. By the same token, evidence of convergent validity for the vocabulary tests can be found via high correlations between test # 3 and test # 4. On the other hand, low correlations between test #1 and test # 3, test # 2 and test # 4 speak to the degree that the tests demonstrate evidence of discriminant validity.

Cronbach and Meehl (1955) argued that construct validation finds importance when no external and outside criterion measure is available for researchers:

> When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the infinite frustration of relating every criterion to some more ultimate standard. Construct validation must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured. (p. 50)

A point which is worthy and Stevenson (1981) makes is very important to set the stage for the research question of the current study. He refers to "elements-by-skills" and also the fact that between them "mutual exclusivity exists" (p. 52).

What are more interesting in Stevenson's paper are the comments he gives about traits and methods. The researcher makes the point of saying that at times it is unclear to say what trait is and what method is. As an example, he refers to interview as a testing device. Simple conversational exchanges between interlocutors can be both a trait and method property. Still as another example he posits the role of oral phenomenon in a test of reading comprehension. In such a test, the main focus is on reading comprehension and not speaking or listening. So, it is one point where the effect of method from trait cannot be distinguished.

Bachman and Palmer (1983) elaborated on two advantages for MTMM designs. First, the paradigm makes it possible for the researcher to examine both convergent and discriminant validity. A second advantage referred to by these scholars is that the designs allow for the researcher "to distinguish the effect of measurement method from the effect of the trait being measured" (p. 156).

2.2 Some Research Studies Using MTMM

Two important studies that deserve attention in this section are that of Bachman and Palmer (1982) and Llosa (2007). In Bachman and Palmer (1982), three distinct traits–linguistic competence, pragmatic competence and socio-linguistic competence were posited as components of communicative competence. They employed a multi-trait multi-method design using four different methods, a writing sample, an oral interview, a multiple choice test and finally a self-rating scale. In their study, grammatical competence includes morphology and syntax. Phonology and graphology were not included because these were considered to be channels as opposed to components. Under the notion of pragmatic competence they

include cohesion, coherence, and vocabulary (surprisingly). The justification for including vocabulary in pragmatic competence as opposed to grammatical competence came from their observation that non-native speakers with a good knowledge of vocabulary were able to carry out quite meaningful pieces of communication as compared to speakers who have a good command of grammatical structures but less knowledge of vocabulary. Finally, under the layer of sociolinguistic competence they included sensitivity to registers, nativeness and registers. To sum up, discourse competence and strategic competence were not included here.

The authors postulate two lines of research paradigms one of which support a general factor for language proficiency. One of the advocates of this paradigm is Oller (1983). Oller's idea of unitary competence hypothesis is pertinent here. Still another line of research views language proficiency as consisting of one general factor and three distinct factors. It is noteworthy to assert that the authors subscribe to the latter view. As a means of determining ways of investigating the construct validation, the researchers used multi-trait multi-method design. According to the authors, MTMM studies two sources of error variance: that due to test method and that due to random measurement error. The authors did employ MTMM and also confirmatory factor analysis. The belief is that in this way trait and method effects or factors are clearly distinguishable. The researchers operated on the assumption that there would be no interaction between the methods and the traits and that the four factors would be uncorrelated with one another. It was found out that one general factor and three correlated trait factors provided barely a significant fit, while a general factor and three uncorrelated factors gave a better image of the relationship among the traits and the methods and explained the nature of communicative competence better. Although the researchers had posited distinctness between pragmatic competence and grammatical competence, the findings did not lend support to the assumption. The two appeared to, in their terms, "cluster

together" (p. 462). On the other hand, a separate sociolinguistic competence factor did emerge. In terms of factor loadings the findings were as follows: all the measures loaded heavily on the general factor except multiple-choice items. Interestingly enough, the largest loadings belonged to using interview and writing methods.

Another research that was conducted within the realm of MTMM designs was the one carried out by Llosa (2007). She attempted to find out if one classroom-based test was valid as compared with a national and standardized test. The pseudonym and the acronym used to show the classroom-based test was LED and the acronym used to describe the criterion measure was CLEDT. The validity of LED was not known and it was to be investigated. The approach adopted was a criterion related analysis. But it is different from the traditional criterion related analysis in that the approach does resort to confirmatory factor analysis as applied to the MTMM design. As discussed above, MTMM designs use more than one trait and more than one method. The traits used in the study were speaking/listening, reading, and writing. Also the methods used here were LED and CLEDT. The design might be shown as in Table 2.

Table 2: Traits and Methods in Llosa (2007)

Traits / Methods	Speaking/ listening	Reading	Writing
LED			
CLEDT			

She arrived at the following conclusions. The CFA MTMM analysis revealed that there were substantial method effects in all grade levels, suggesting that the scores on the LED Classroom Assessment do not just reflect trait, but also reflect factors associated with the testing instrument as well.

It was observed that MTMM designs are classical but useful ways of investigating the construct validity of a test. In the University of Tehran English Proficiency Test, there are at least two traits and two methods. So, the conditions are ripe for the MTMM design to be employed.

The present study intends to measure the degree of correlation among traits being tested and to see, on the one hand, whether it changes when the traits are same or different, and, on the other hand, whether it varies when methods of testing are same or different. In other words, the test is deemed valid if different traits yield a low correlation using or not using the same methods. It is also true that a high correlation arrived at by correlating the same traits using or not using the same methods render the test valid.

3. Methodology

3.1 Participants

The participants were 3,398 PhD candidates chosen from the total population of 8694 testees who took the University of Tehran English Proficiency Test in Esfand of 1385 (February 2007). The details of the test will be given in the subsequent sub-section. The researchers discarded the outliers. In this study, outliers were operationally defined as those testees having scored well above the mean or well below the mean. After outlier analysis was done, in order to avoid too similar subjects in terms of performance, we omitted participants half a standard deviation above and below the mean. The participants belonged to different fields of study, e.g., physics, chemistry, etc.

3.2 Instrumentation

The University of Tehran English Proficiency Test (the UTEPT) is a three-section test consisting of 100 items which are grammar, vocabulary and reading comprehension. The grammar section has 35 items. The first 20 items are multiple choice completion items. The second 15 items are error identification.10 items deal with grammar and vocabulary tested in context. The next section deals with vocabulary. The section comes into two parts. Part one has 10 items. Part two has 10 items. The last section is concerned with reading comprehension. This section consists of six passages with 35 Multiple-choice items.

3.3 Data Analysis

In the MTMM design that was of interest in the present study, a correlational tool was employed. This design uses more than one trait and more than one method. The traits used were grammar and vocabulary. And the methods used were multiple choice as well as contextualization techniques. Both convergence and divergence were tested in the data. In other words, a high correlation was to be expected between the same traits using different methods. Likewise, a low correlation was expected between different traits using the same methods.

It is to be noted that the reading comprehension section was not exposed to investigation via the MTMM design because this trait was tested through one single method and the design was not appropriate for this trait.

4. Results and Findings

4.1 MTMM Designs as Pieces of Evidence for Convergent and Divergent Validity

As for the research question of the present study which inquires about the correlation between traits being tested, same or different, and the methods of testing, different or same and following Palmer and Groot (1981), the researchers came up with table 3. As traits, grammar and vocabulary were considered. As methods, two methods of testing the two traits

were considered. These methods are multiple choice completion and contextualization. It is to be remembered that contextualization might safely be called a cloze test. The researchers also did follow Campbell and Fiske (1959:81) who favor using more than one trait as well as more than one method. In this study there are two traits and two methods. So, the conditions are met for the application of this design. The same trait might be tested in more than two ways or more. Or there might be more than one trait involved. But as reiterated, the current study used the minimum number of traits, i.e., two and the minimum number of methods, i.e., two. But in the case of monotrait multimethod designs, more than two methods were included.

Table 3: Traits and Methods in the Study

Methods / Traits	Discrete -Point Multiple Choice	Contextualized Multiple Choice
Grammar	Test 1	Test 2
Vocabulary	Test 3	Test 4

Bachman and Palmer (1983) convince the readers that appropriate reliability indices are essential for any convergent/discriminant validity study. Following them, the reliability indices for the measures were estimated. The results are presented in table 4.

Table 4: Reliability Statistics for Different Sections of the Test

Traits and Methods / Reliability Method	DPMC		CMC	
	grammar	vocabulary	grammar	vocabulary
Alpha	.60	.35	.50	.40
Split half	.50	.25	.35	.20

DPMC= Discrete point multiple choice completion
CMC=cloze multiple choice

Building on Clifford (1981), the researchers came up with an MTMM convergent and divergent validation matrix as shown in table 5. It makes it possible to investigate the correlation among the same trait tested through different methods and the different traits tested through the same methods. In the case of the former, a high correlation is expected. In the case of the latter, a lower correlation is to be anticipated. One can test if those predictions are actually borne out.

Table 5: A Convergent/Divergent Matrix

	DPMC vocabulary	DPMC grammar	CMC vocabulary	CMC grammar
DPMC vocabulary	1	.07	.50	.04
DPMC grammar	--	1	.06	.48
CMC vocabulary	--	--	1	.03
CMC grammar	--	--	--	1

Al correlations are significant at .01 level.

Based on table 1, our expectation is that a high correlation will be produced between DPMC vocabulary and CMC vocabulary on the one hand and DPMC grammar and CMC grammar on the other in the matrix because the same trait is

tested by using different methods. As it can be observed, in the case of the former the correlation is .50 and for the latter the correlation is .48. In both cases, evidence has been provided for the convergent validity of the test under investigation.

In order to demonstrate if the test has exhibited discriminant validity, DPMC vocabulary and DPMC grammar are expected to produce a low correlation. The correlation between them is .07 which is low enough to maintain that two similar methods tapping different traits show acceptable level of discrimination. By the same token, a low correlation was anticipated between DPMC grammar and CMC vocabulary. They are different methods to tap different traits. It was revealed be .06. Again this is low enough to provide evidence for the discriminant validity of the test. The correlation obtained from scores from DPMC vocabulary and CMC grammar is .04. This was expected because again different traits were tested through different methods. Finally, the same methods taping different traits are supposed to yield low correlation. The prediction is borne out. The correlation produced is .03.

4.2 Monotrait Multimethod Designs as Evidence of Convergent Validity

According to Clifford (1981), convergent validation can be estimated by computing the correlation between the test and another independently developed test of the same trait. In the proficiency test under examination, the trait "vocabulary" was tested by four methods, discrete-point multiple choice paraphrase, discrete-point multiple choice completion, vocabulary in context and finally by inserting vocabualry in the reading comprehension section. The results are shown in table 6.

Table 6: Monotrait-Multimethod Matrix for Vocabulary

	DPMCP	DPMCC	VC	VIIRC
DPMCP	1	--	--	--
DPMCC	.09	1	--	--
VC	.09	.01	1	--
VIIRC	.25	.09	.07	1

All correlations are significant at .01 level.

DPMCP=discrete point multiple choice paraphrase
DPMCC =discrete point multiple choice completion
VC= vocabulary in context
VIIRC=vocabulary inserted in the reading comprehension

The matrix above illustrates the monotrait multimethod design (Bachman & Palmer, 1981). The vocabulary trait has been tested by four different methods. The researchers expected a high correlation to emerge. But it did not happen. But the correlations are congruent with what Bachman and Palmer (1981) assert; "Monotrait-heteromethod correlations should be significantly higher than zero" (p. 158). The correlations are indeed higher than zero and significant (except the correlation between VC and DPMCC which is low and non-significant at .01 level.

The highest correlation belongs to the one between the vocabulary tested in the reading comprehension and vocabulary tested through multiple-choice paraphrase. This is not surprising because in both cases vocabulary items appear in context. The lowest correlation went to the one between discrete-point multiple choice completion and vocabulary in context.

By the same token, grammar was tested using three methods. One was multiple choice completion, the second method was error identification (here referred to as written expression). Finally, the third method was grammar in context.

The acronyms used are MCC, EI, and GIC respectively. The evidence for convergent validity can be found in table 7:

Table 7: Montrait-Multimethod for Grammar

	Multiple Choice Completion	Error Identification	Grammar in Context
Multiple Choice Completion	1	--	--
Error Identification	.27	1	--
Grammar in Context	.11	.29	1

All correlations are significant at .01 level.

The highest correlation belongs to the one between grammar in context and error identification. The lowest correlation is between grammar in context and multiple- choice technique. The moderate correlation is between error identification and multiple-choice technique.

5. Conclusions and Discussions

Table 5 has managed to give us evidence of the construct validity of the test through convergent and divergent analyses. Based on table 4, our expectation is that a high correlation will be produced between test 1 and test 2 as shown in the matrix because they test the same trait by using different methods.

In table 6, low, albeit significant, correlations speak to the effect exhibited by the method factor. The lowest correlation index is tantamount to the greatest effect. By the same token, the highest correlation index is equal to the least effect as exercised by the method factor. The method factor is greatly at work when the trait of vocabulary is tested through multiple choice completion and contextualization. On the other hand, the effect of method is the least when the trait of vocabulary is tested through being inserted in the reading passage and multiple choice paraphrase items. This is convincing because

the methods share commonalities. In other words in both cases, the vocabulary items are underlined and the synonyms are sought. In all cases, the effect of method factor is quite evident but is almost the same across the measures.

In table 7, low correlations obtained among measures tapping the same traits is a piece of evidence provided for the effect of the method factor. The method factor is quite apparent in the case of grammar tested through multiple choice completion and contextualization. However, method factor does play almost an equal role when grammar is tested through error identification and multiple-choice completion, and through error identification and contextualization.

It was observed that for the vocabulary and grammar sections, the MTMM design was a good way of investigating validity. The application of this procedure showed that the test was valid. Specifically, it was shown that the same traits tested through the same methods had a higher correlation than that of the same traits having been tested through different methods.

It must be mentioned that this design was very much appropriate in the light of reliability indices which were obtained in the sub-tests. Had the reliability indices been even higher, a better picture of validation would have been provided by the design.

To sum up, using the MTMM design was very helpful in collecting evidence for the validity of the test.

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* New York: Cambridge University Press.

Bachman, L., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, J. D. Groot, & G. Tropser. (Eds.), The construct *validation of tests of communicative competence, including proceedings of a colloquium at TESOL '79*, Boston February 27- 28, 1979.

Bachman, L., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16,* 449-465.

Bachman, L., & Palmer, A. S. (1983). The construct validity of FSI oral interview.In W.J. Oller, (Ed.), *Issues in langauge testing research (*pp. 154-169). Rowely, MA: Newbury House.

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi -trait multi- method matrix. *Psychological Bulletin, 56,* 81-105.

Clifford, R. T. (1981). Convergent and discrimination validation of integrated and unitary language skills: The need for a research model. In A. S. Palmer, J. D. Groot, & G. Tropser. (Eds.), *The construct validation of tests of communicative competence, including proceedings of a colloquium* at TESOL '79, Boston February 27- 28, 1979.

Cronbach, L. J., & Meehl, P. E. (1955). Constrcut validity in psychologicl tests. *Psychological Bulletin, 52,* 281-302.

Llosa, L. (2007). Validating a standards-based assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24,* 489-515.

Long, S. J. (1983). Confirmatory factory analysis: *A preface to LISEREL.* Sage University series on Quantitative Applications in the Social Sciencesm series no. 07-033. Beverly Hils, Callifornia: Sage.

Messick, (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In Wainer, H. and Braun, H. *Test validity*. Hillsdale, NJ: Erbaum, 33-45.

Oller, W. (1983). Evidence for general language proficiency factor: An expectancy grammar. In W. J. Oller. (Ed.), *Issues in langauge testing research* (pp. 3-10). Rowely, MA: Newbury House.

Roever, C. (2001).Web-based language testing. *Language Learning and Technology,* 5(2), 84-94.

Palmer, A. S., & Groot, P. J. M. (1981). An introduction. In A. S. Palmer, J. D. Groot, & G. Tropser. (Eds.), *The construct validation of tests of communicative competence, including proceedings of a colloquium at TESOL '79, Boston February* 27- 28, 1979.

Stevenson, D. K. (1981). Beyond faith and face validity: The multitrait- multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A. S. Palmer, J. D. Groot, & G. Tropser. (Eds.), *The construct validation of tests of communicative competence, including proceedings of a colloquium at TESOL '79*, Boston February 27-28, 1979.

Zand Karimi, F. (2005). *Investigating construct validity of the sub-skills of the reading section of University of Tehran English Proficiency Test.* Unpublished MA Thesis: University of Tehran.