The Construct Validity of a Test: A Triangulation of Approaches

MOHAMMAD SALEHI
Sharif University of Technology

Bio Data:
Mohammad Salehi is currently a faculty member of the Languages & Linguistics center at Sharif University of Technology and holds a PhD degree in TEFL from the University of Tehran. He got his MA degree from the University of Allameh Tabatabai in TEFL. He earned a BA degree in English literature from Shiraz University. He has taught in University of Kashan, University of Teacher Training, University of Amirkabir, Azad University of Karaj, University of Tehran, and University of applied sciences. He has presented articles in Turkey, Cyprus, Dubai, Armenia, Australia, Italy and Iran. His research interests include language testing and second language acquisition research. He has also written books on language testing and vocabulary.

Abstract
Three approaches of validation enquiry were applied on the data obtained from a proficiency test carried out with 3,398 PhD candidates as a partial requirement for entering PhD program in Iran. The data obtained from the reading section were subjected to an exploratory factor analysis (EFA). The EFA yielded nine factors in the reading section. Furthermore, a design of multitrait-multimethod was also investigated, the traits being grammar and vocabulary and the methods, multiple choice and cloze test. The test demonstrated both convergent and discriminant validity. A third approach was intersub-test correlations as elaborated on by Alderson et al. (1995). For this, different sub-parts of the instrument were correlated with one another and with the total test. The test also showed construct validity as investigated via this approach. It can be concluded that the test had construct validity as revealed by these three approaches.

Keywords: Exploratory factor analysis, Multi -trait multi-method, Varimax rotation, Factor structure, Construct validity

Introduction

Validation is an important enterprise especially when the test is a high stakes one. Admission tests for universities or other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations (Roever, 2001). According to Messick (1988), if the validity of a test is not known it might have

undesirable consequences for the society at large. University of Tehran English Proficiency Test (the UTEPT) qualifies as a high stakes test as almost ten thousands test takers take it on a yearly basis. It also has life changing implications for the test takers involved.

Although studies in language testing have used single lines of enquiry in test validation (LLosa, 2007, Bachman and Palmer, 1983), this study uses triangulation in test validation and enquiry. The value of using a triangulation approach is clear from Anderson, Bachman, Perkins, and Cohen's (1991) statements in favor of it.

Perhaps the greatest insight gained from this investigation is that more than one source of data needs to be used in determining the sources of reading comprehension test items. By combining sources of data such as those examined in this study (i.e., data from readers' retrospective think-aloud protocols, test content evaluation, as well as the traditional test performance statistics) greater insights are gained into the reading comprehension process as well as the test taking process. (p. 61)

So in this study attempts will be made to bring data from multiple sources. Although a number of studies have already been conducted in the area of language testing (e.g., Bachman & Palmer, 1983), almost no study, to the knowledge of the present researcher, has brought pieces of evidence from different sources to bear on language skills. The uniqueness of the present study lies in the fact that it deals with the concept of multiple perspectives in the validation inquiry. The UTEP had not been exposed to vigorous investigation before. Given the fact hat it is a high stakes test, a study is on order.

Review of the Related Literature

Approaches to Construct Validation

There have been several approaches to test validation. A sketch of the approach of Alderson, Clapham, and Wall (1995) approaches to validation is the most appealing. The first approach that they mention is the correspondence with theory. In other words, the test results are supposed to confirm the theory. The authors remind us that the theory itself is not called into question. The second approach they mention 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 mention 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 correlation. 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 touch upon is factor analysis which will be explained in the following sections. Another approach is multitrait- multimethod approach which will be elaborated on in due course. Finally, the last approach is taking account test bias and actually assessing the role of background knowledge, gender, race, etc.

Factor analysis. Baker (1989) maintains that "factorial analysis is broadly speaking, to simplify a variety of sets of scores (which we will call variables) for a given population" (p. 62).

There are two major types of factor analysis: exploratory and confirmatory. As to exploratory factor analysis, Bachman (1990) says, "In the exploratory mode, we attempt to identify the abilities, or traits that influence performance on tests by examining the correlations among a set of measures" (p. 260). Bachman (1990) offers the following insight about confirmatory factor analysis: "In the confirmatory mode, we begin with hypotheses about traits and how they are related to each other, and attempt to either confirm or reject these hypotheses by examining the observed correlations" (p. 260).

Multitrait-Multimethod designs. Perhaps the pioneers for Multitrait Multimethod designs are Campbell and Fiske (1959). Palmer and Groot (1981) have us to believe 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 followed by research studies.

Test scores may be the function of trait and the method used to test it. For example, grammar 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 multitrait- multimethod designs are the notions of convergent and divergent validity.

First convergent validity is elaborated on. 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 correction is supposed to be high because in each case we are testing grammar and any difference can be attributed to the effect of the method.

Divergent or discriminant validity is logically related to the convergence of scores. Lets' illustrate the difference between convergence and divergence through an example. Vocabulary and grammar are expected 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 are 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 that the two skills tap similar skills like vocabulary knowledge and world knowledge.

Protocol analysis. Perhaps the first scholar to draw our attention to the feasibility of gathering evidence through verbal reports is Cohen (1984). Cohen argues that a mismatch between the intentions of test makers and the thought processes of testees will call into question the validity of a test. In other words, if an inference item is conceived to be a reference one by testees, this is a blow to the

validity of a test. Kormos (1998) clarifies the difference between think-alouds, introspection and retrospection. For think alouds, researchers instruct the subjects to verbalize whatever that occurs to them while performing a task. For introspection, the subjects are not only asked to verbalize but also to justify their thought processes. Finally, retrospection is different in that the subjects are supposed to verbalize after they have performed the task. According to Kormos, (1998) the disadvantage is that the subjects need to transfer information from the long term memory to the short term memory which can jeopardize the accuracy of the verbalization.

Camps (2003) maintains that recent studies have shown the usefulness of think-aloud protocols in understanding learners' cognitive processes as they perform tasks designed to help them make form-meaning connections when processing input.

Research Questions:

1. Do the test items in the 'Reading Comprehension' sections of the UTEPT distinctly measure various sub-skills?
2. Is there any correlation between traits being tested and the methods of testing?
3. What is the degree of correlation among sub-parts of the test?

Methodology

Participants

The participants included in the present study are 3,398 testees truncated from the total population of 8,696 testees who took the UTEPT in February 2007. Outliers were discarded. The participants majored in different fields of study, including physics, chemistry, theology, etc. They took the test as a partial requirement for entering the PhD programs.

The Instrumentation

The UTEPT. The test consists of 100 items. The three sections of the test 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 (items 36 to 45) deal with grammar and vocabulary tested in context. The next section deals with vocabulary. This section is divided into two parts; part one has 10 items (items 46 to 55) and part two has 10 items (items 56 to 65). The last section is concerned with reading comprehension. This section has 35 items consisting of six passages.

Data Collection

The data were collected in two stages. In the first phase of the study, the data obtained from a total of 8,696 testees on the UTEPT were analyzed and the scores of 3,398 testees were selected for the final analysis. The value of the data collected is quite apparent because the respondents did the test with motivation as their admission into the PhD programs hinged on the outcomes of the test. As a matter of

fact, certain faculties would set the results of this exam as a precondition for technical exams. The researcher had no control over the data collection.

Data Analysis

Exploratory factor analysis using Principal Components Analysis (PCA) was used. Because the data set was large enough to conclude that the distinction between PCA and other types of factor analysis was insignificant (Kline, 1944). Also factor loadings below .30 were ignored (Hatch and Lazaraton, 1991). Furthermore, correlational analyses were used.

Results

Factor Analysis (answer to the first research question)

To answer the first research question, exploratory factor analysis with using principal components analysis with varimax rotation was employed. The results are shown in Table 1.

As it can be seen, eleven factors were extracted. Factor loadings below .30 were ignored and they cannot be seen in the Table. Factor impureness is noticeable. In other words, there are items that load on more than one factor. In case of these items the variance is shared among factors and so not very high loadings are observed. Items 91 and 92 are examples of factor pure items and items 71 and 88 are examples of factor impure items.

Table 1
Extracted Factors on Reading Comprehension Items by Principal Components Analysis

	Component										
	1	2	3	4	5	6	7	8	9	10	11
q066				.589							
q067				.582							
q068				.388							
q069				.451				-.361			
q070			-.395		-.339			.544			
q071											
Q072	.456										
Q073								.606			
Q074							.365	.432			
Q075										.504	
Q076						-.614					
Q077										.592	
Q078					.388	.694					.449
Q079						.366					
Q080										-.511	-.449
Q081		.316									
Q082									-712.		
Q083	.391								-.334		
Q084									.464		
Q085					.514						
Q086			.327								
Q087											
Q088	.552	.562									
Q089		.599									.692
Q090	.624										
Q091											
Q092							.627				
Q093					.585						
Q094	.495	.322	.462								
Q095			.386				.322				
Q096			.599				.388				
Q097					.339						
Q098			-.434				-.493				
Q099								-.307			
Q100		.440									

The extracted factors and their explanations appear below:

Factor one. Items 72, 83, 86, 90 and 94 loaded on factor one. These items appear below:

72. The word "acquire" in line 6 is closest in meaning to --------.
A. occupied C. organized
B. obtained D. operated

83. The word "excite" in line 12 is closest in meaning to -------.

A .stimulate B. stick C. strike D. summon

86. The word "tangible" in line 4 is closest to -----.
A. visible B. verified C. various D. violent

90. The word "vast" in line 3 is closest in meaning to….
A. large B. basic C. new D. urban

94. The word "potential" in line 16 is closest in meaning to --------
A. certain C. popular
B. improved D. possible

As it can be seen, this is clearly a vocabulary factor and all the five items are vocabulary items tested within the reading passages.

Factor two. Items 81, 88, 89, 94 and 100 loaded on this factor. These items are shown below:

81. According to the passage, investments in service are comparable to investments in production and distribution in terms of the ------.
A. insufficient analysis that managers devote to them
B. tangibility of the benefits that they tend to confer
C. increased revenues that they ultimately produce
D. basis on which they have to be weighed

88. With which of the following subjects is the passage mainly concerned?
A. Types of mass transportation
B. Instability of urban life
C. How supply and demand determine land use
D. The effects of mass transportation on urban life

89. The author mentions all of the following as effects of mass transportation on cities
EXCEPT …
A. growth in city area B. separation of commercial and residential districts
C. changes in life in the inner city D. increasing standards of living

94. The word "potential" in line 16 is closest in meaning to --------
A. certain C. popular B. improved D. possible

100. Which of the following best organizes the main topics addressed in this passage?
A. I. Entertainment in small-town, nineteenth-century America
II. How medical tricksters take advantage of common fears
B.I. Products distributed by nineteenth century travelling medicine shows
II. Influence of travelling medicine shows on modern-day medical trickery
C. I.-Nineteenth-century medical sales techniques
II.-Contemporary forms of medical trickery

D. popularity of nineteenth century traveling medicine shows
II. how to guard against modern-day medical trickery

An analysis of the items under this factor shows that it is mostly a main idea factor. Item 94 looks like a non-belonging one. The reason can be attributed to the fact that the item is not factor pure and is also related to factor one.

Factor three. Items 71, 87, 95, 96, 97, and 99 loaded on this factor. Items 71, 87, 96 and 99 have the lowest factor loadings. Item 97 has the highest factor loading. Finally, item 95 has a low factor loading. Item 71 is not factor pure and also loads on factor 5 extracted and to be elaborated on later. Item 87 is a factor- pure item. But the point is that it does not have a high factor loading. Item 95 is not factor pure and shares variance with factor 7. But at the same time, it has a moderate factor loading. Turing to item 96, it has a low factor loading and shares variance with factor 7 in the same way as the preceding item did. The next item to be discussed is item 97, which has the largest factor loading of all the variables which are items in this study. The last item under this factor is item 99. The item is low in factor loading and is not factor pure. It is time we turned to the items briefly discussed in terms of factor loadings and see whether factor naming can be done.

71. The word "boasted" in line 1 is closest in meaning to --------.
A. possessed B. promised C. provided D. proposed

87. The discussion of the regional bank serves which of the following functions within the passage as a whole?
A. It describes an exceptional case in which investment in service actually failed to produce a competitive advantage.
B. It demonstrates the kind of analysis that managers apply when they choose one kind of service investment over another.
C. It provides an example of the point about investment in service made in the first paragraph .
D. It illustrates the pitfalls of choosing to invest in service at a time when investment is needed more urgently in another area.

95. According to the passage, what was one disadvantage of residential expansion?
A. It was expensive. B. It happened too slowly.
C. It was unplanned. D. It created a demand for public transportation.

96. The author mentions Chicago in the second paragraph as an example of a city
A. that is large
B. that is used as a model for land development
C. where land development exceeded population growth
D. with an excellent mass transportation system

97. Which of the following can be the best title of the passage?
A. Medical entertainment B. Common practice treatment

C. Medical treatment business D. Traveling shows

99. Which of the following is the best meaning of the word **quacks** as it is used in the second paragraph of the passage?
A. health care organizations B. medical supply companies
C. traveling entertainers D. dishonest medical practitioners

One might refer to this factor as one related to inference. There are a few points that need to be made about the factor. First, items 99 and 71 have been loaded on this factor. This is surprising because they are vocabulary items and the expectation was that they would be loaded on the first extracted factor. The second point is the one that should be made about item 97. This item has, as mentioned before, the largest factor loading of all the items collected under the factor. This item has one peculiar characteristic: it taps topic identification which is an endeavor in inferencing.

Factor four. Items 66, 67, 68, 69 loaded on this factor. The factor loadings are relatively high .589, .582, .388, and .451, respectively. Let's have a look at the items and name the factor:
66. It is pointed out in the passage that traditionally animals are believed to------.
A. imitate man in many ways B. behave instinctively and logically
C. have comparable intelligence D. act on instinct

67. According to the passage modern research suggests researchers to consider -------.
A. why animals behave differently under different circumstances
B. the possibility of intelligence in animals
C. the improvement of animal behavior
D. how animals can be made to acquire new skills

68. According to the passage in the light of modern research, our traditional assumption about animals' behavior ------.
A. have been totally disproved B. were based on scientific fact
C. have been reconsidered D. should never have been questioned

69. The word "startling" in line 5 is closest in meaning to -------.
A. amusing B. appealing C. activating D. astonishing

A close inspection of the items reveals that they are directly- stated question items. All four items are based on a single passage. These items are often easy items. As a matter of fact, relating the performance of the testees to these items confirms the claim. The facility values for the mentioned items are: .61, .76.5, .61, and .33, respectively. Except for item 69, other items are considered to be relatively easy. These items are so incongruent with the rest of the items. It looks that the test assembler was not equipped with a table of specifications.

Factor five. Items 71, 79, 85, 93, and 98 loaded on this factor. Item 71 is not factor pure and loads on factor 3 as much as it does on this factor. Item 79 is not

factor pure either and it loads more on factor 11 than it does on this particular factor. Item 85 has a relatively high factor loading and is factor pure. In the same token, item 93 is factor pure and has a factor loading close to that of item 85. Our expectation is that this factor, whatever it is, is going to be related to these two items. The last item is not factor pure and it cannot be expected to contribute to this factor. The above mentioned items are shown below:

71. The word "boasted" in line 2 is closest in meaning to --------.
A. possessed B. promised C. provided D. proposed

79. The term "information society" emphasizes ------.
A. the social nature of knowledge B. popular knowledge
C. social convention D. post industrial society

85. The word "merit" in line 15 is closest to -----.
A. aspect B. action C. advantage D. attest

93. Why does the author mention both Boston and Chicago?
A. To demonstrate positive and negative effects of growth
B. To show that mass transmit changed many cities
C. To exemplify cities with and without mass transportation
D. To contrast their rates of growth

98. Which sentence, if inserted into the blank line in the second paragraph, would be most consistent with the writer's purpose and intended audience?
A. I think you should at least make an effort to determine who prepared the report and how the researchers arrived at their conclusions.
B. They need to ask questions about who conducted the research and what testing procedures were used.
C. They must comprehensively probe the fitness of researchers and incisively evaluate the sufficiency of their methodology.
D. You should try to learn something about who did the research and how they did it.

The items did not turn out to behave in the manner they were expected to. Item 85 is a vocabulary item. Item 93 is not a vocabulary item; it is more related to reasoning ability than it is to simple vocabulary knowledge.

Factor six. Items 76, 78, and 79 came to be loaded on this factor. Item 76 is factor pure with negative factor loading. Item 78 is factor pure with a high factor loading. Lastly, item 79 is not factor pure and also loads on factor 11. So, probably we are going to count on items 76 and 78 to help us in factor naming. First, items should be looked at:

76. Higher education furnishes the graduates primarily with ------.
A. profession B. discipline C. knowledge D. service

78. The word "what" in line 5 refers to -------.
A. application B. context of education
C. program of universities D. content and methods of certain subjects

79. The term "information society" emphasizes ------
A. the social nature of knowledge B. popular knowledge
C. social convention D. post industrial society

Item 78 has the highest factor loading and it is a reference item. Probably, all items are concerned with word paraphrase.

Factor seven. Items 74, 92, 95, 96, and 98 loaded on this factor. The items can be analyzed in terms of factor pureness. Item 74 is not factor pure; it also shares variance with factor 8. It loads more on factor 8 than it does on this factor (i.e., factor seven). So, not much investment can be made on the contribution of this factor. Item 92 has the highest factor loading of all the variables (here items). Also, it is a factor pure item. This item has made the greatest contribution to the factor. Items 95 and 96 loaded on this factor as they did on factor three. Finally, item 98 loaded on this factor as it did on factor 5. So, emphasis needs to be placed on item 92 to help us to come up with a name for the factor.

92. The word "many" in line 18 refers to
A. people B. lots C. years D. developers

It comes as no surprise that this item has the largest factor loading of all as well as being a pure-factor item. The reason is that this item tests a grammatical point in the language; no other item in the section behaves similarly.

Factor eight. Items 69, 70, 73, 74, and 99 came to be included under this factor. Item 69 is not factor pure and it also loaded on another factor. As a matter of fact, the impureness of this in terms of factor loading is evident in the fact that the item is incongruent with the set of other items belonging to directly stated questions. Apart from that item, one should see what has happened to item 70. This item has a large, although not the largest, factor loading. The factor is probably looking for a great contribution from the item. Next, there is item 73 with the largest factor loading of all the items and expected to make a great contribution to the factor which has been extracted. The last two items are not factor pure and are not expected to be of any help in naming the factor. The two most contributing items, i.e., items 70 and 73 are shown below:

70. According to the passage, in the early years of universities -------.
A. most students wanted to train for a profession
B. medicine was the most popular subject for study
C. the church disapproved of much of their teaching
D. the majority of students came from upper class families

73. According to the passage, since most of the early universities enjoyed the support of the church, ----
A. the number of students they admitted increased rapidly

B. state authorities granted then various rights
C. law naturally became one of the major subjects offered
D. the education offered was free of charge

The two items have appeared under the same factor for very good reasons. One is that they are both based on the same passage. But more important than that is the fact that the items fall somewhere between inference and main idea types which place a lot of demands on the test taker and directly stated questions which are not as demanding for the test takers. So, this factor can be called "understanding through paraphrase."

Factor nine. Items 82, 83, and 84 loaded on this factor. Item 82 is factor pure. Item 83 is not and it also loads on factor one. So, this item is probably going to be a vocabulary factor. Finally, item 84 is also factor pure and responsible for accounting for variance. Let's scrutinize items 82 and 84 and see if our prediction about the characteristic of item 83 is borne out.

82. The passage suggests which of the following about service provided by the regional bank prior to its investment in enhancing that service?
A. It enabled the bank to retain customers at an acceptable rate.
B. It threatened to weaken the bank's competitive position with respect to other regional banks.
C. It had already been improved after having caused damage to the bank's reputation in the past.
D. It was slightly superior to that of the bank's regional competitors

83. The word “excite” in line 14 is closest in meaning to -------.
A. stimulate B. stick C. strike D. summon

84. The passage suggests which of the following about service provided by the regional bank prior to its investment in enhancing that service.
A. It threatened to weaken the bank's competitive position with respect to other regional banks.
B. It enabled the bank to retain customers at an acceptable rate.
C. It had already been improved after having caused damage to the bank's reputation in the past.
D. It was slightly superior to that of the bank's regional competitors.

Turning to our prediction about item 83, it can be seen that it was borne out. But as for items 82 and 84, it becomes evident that both use the word "suggest" in their stems leading us to conclude that the concern of the items is to tap "drawing conclusions."

Factor ten. Items 75, 77 and 80 came to be loaded under this factor. Items 75 and 79 are factor pure and are likely to be accountable for the greatest contribution to the factor as opposed to item 80 which does not load on a single factor; it also

loads on factor 5. The two items should be inspected to see if they can tell us anything about the factor:

75. Which of the following can be the title for this passage?
A. Knowledge and civilization B. Educational knowledge
C. knowledge in higher education D. Crucial role of knowledge

77. Higher education furnishes the graduates primarily with ------.
A. profession B. discipline C. knowledge D. service

Both items are based on the same passage. Item 75 is looking for an identification of a title for the passage. Item 77 is indirectly having the same function. It is worthy of note that in both items, the correct answer has the word "knowledge" in them. As a matter of fact, some kind of manipulation of the items leads us to the conclusion that the items have similar traits. In item 75, the key phrase is "knowledge in the higher education." Now, in item 77, we can combine the stem with the correct choice and come up with the same proposition. In other words, "higher education furnishes the graduates with knowledge" is propositionally the same as "knowledge in higher education."

Factor eleven. Items 79, 81 and 91 loaded on this factor. The first two items are not factor pure and item 91 is held accountable for explaining the variance. Item 91 is shown below:

91. The word "sparked" in line 11 is closest in meaning to -----.
A. brought about B. surrounded C. sent out D. followed
Item 91 has surprisingly loaded on this factor. It is the point where factor analysis should be combined with logic.

Multi-trait Multimethod (Answer to the second research question)
To answer the second research question and following Palmer and Groot (1981), the researcher came up with Table 2. As traits, grammar and vocabulary were considered. As methods, two methods of testing the two traits were considered. The researcher also did follow Campbell and Fiske (1959) who favor using, "more than one trait as well as more than one method" (p. 81).

Table 2
Multitrait Multimethod Design

Methods Traits	Discrete -point Multiple Choice	Contextualized multiple choice
Grammar	**Test 1**	**Test 2**
Vocabulary	**Test 3**	**Test 4**

The results of the correlations are shown in Table 3. As it can be seen the correlations are low and this speaks to construct validity of the test under study.

Table 3
Convergent/Divergent Correlations

Method	Trait				
DPMC	Grammar	1			
DPMC	Vocabulary	.07	1		
CMC	Grammar	.48	.04	1	
CMC	Vocabulary	.06	.50	.06	1
		DPMC	DPMC	CMC	CMC
		Gr.	Voc.	Gr.	Voc.

DPMC=discrete point multiple choice
CMC=cloze multiple choice

Intersubset Correlations (Answer to the third research question)
To answer the third research question and following Alderson et al. (1995), the researcher came up with the following correlation matrix as shown in Table 5. In the present study, sub-parts consisted of grammar, vocabulary, and reading comprehension. Very much akin to the study quoted in Alderson et al. (1995), the correlation of each sub-part with the total test and the total minus the subpart itself were computed.

Table 4
Inter- Subtest Correlation Matrix

	Grammar	Vocabulary	Reading	Total	Total minus self
Grammar	1	.07	.10	.74	.11
Vocabulary	.07	1	.19	.54	.16
Reading	.10	.19	1	.61	.18
Total	.74	.54	.61	1	1

N=3,398

Alderson et al. (1995) remind us that a low correlation within the range of .30 to .50 is a substantial piece of evidence for the distinctness of traits. As it can be seen in Table 5, the correlations here support the separateness of the traits, namely grammar, vocabulary and reading comprehension. They range from .07 (the lowest) to .19 (the highest). The contribution of each trait to the overall test score can be computed by correlating the results of a trait with the total score. It can be seen that the vocabulary contributes little to the overall picture. The highest correlation belongs to that of grammar. The next is reading. What Alderson et al. argue is that a correlation of each trait with the total test including the trait itself would push up the correlation index as it does in this study. To solve this problem, it is recommended to correlate the trait with the total test excluding the subtest. This was followed in the present study and the indices dropped considerably. An interesting observation was

made. This time, the highest correlation belongs to that of reading and total test minus reading. It was expected that the indices in the final column would follow the same pattern as the preceding column. This runs counter to the study reported in Alderson et al. The study is that of Alderson et al. (1986). Let's have a look at the results of the study in the form of a table, Table 7, as reported in Alderson et al. (1995).

Table 5
Inter- Subtest Correlation Matrix Reported in Alderson et al. (1995)

	Reading	Proficiency	Writing	Oral	Total	Total Minus Self
Reading	-	.53	.27	.44	.73	.50
Proficiency	.53	-	.43	.66	.84	.72
Writing	.27	.43	-	.45	.66	.46
Oral	.44	.66	.45	-	.86	.66
Total	.73	.84	.66	.86	-	-

As the findings of the table show, the correlations in the last column dropped to a level which is consistent in terms of size with those of the one column preceding it.

Discussion

As it was shown in Table 1, the reading section of the test has clear factor structure. This provides evidence for the construct validity of the test. One problem was overfacoring by which it is meant that the factors are more than expected. There are just 35 reading comprehension items which lent themselves to 11 factors. This has to be accounted for. One explanation can be that the 35 items belong to different paradigms in language testing. The test maker must have opted for ILTES, TOEFL, FCE, etc. The other problem was that some factors were represented only by one item. The reason can be that this item taps only one construct in a way that no other item does. The item may have been taken from somewhere without it being in harmony with the rest of the items. One other oddity was the fact a vocabulary item was collected under factor eleven. It is the one of the cases where one must apply logic and not rely on factor analysis machine (Preacher and MacCallum, 2003).

As Table 3 showed, the high correlation (r=.48) between CMC and DPMC is indicative of the convergent validity of the measure as the same trait is being tapped albeit using different methods. This is to ensure that method does not induce error into the process (Bachman, 1990). In a similar vein, the high correlation (r=.50) between CMC vocabulary and DPMC vocabulary is not surprising because the same trait is being measured via different methods. Again, this provides evidence for the convergent validity of the test. Apparently the test is free from method error. On the other hand, the correlatonal indices are not high enough. Higher correlational indices would have provided more solid pieces of evidence. Turning to divergent validity, the correlational indices are low enough (r=.04, r=.06, and r=.07) to warrant distinctness between the traits.

The results of Table 5 shed light on intersubtest correlation as another line of enquiry for test validation. The highest correlation belonged to that of correlation followed by reading and vocabulary. This was strange. It cannot be the case that the performance on the UTEP was largely related to grammar. As suggested by Alderson et al, the trait itself was excluded in the correlational analysis. By so doing, the correlatinal indices changed. This time the most contributing trait was reading followed by vocabulary and grammar. This is quite fathomable on the grounds that success on the UTPET is largely a function of the performance on the reading section. A consequential validity study can furthermore reveal that reading ability is a good predictor of academic success when they actually enter the university. There happens to be a discrepancy between the results of descriptive statistics and those of the inferential statistics (here correlation). Table 5 showed that grammar had the highest correlation of all with the total test. This means that it had the most contribution to the overall picture. This interestingly is supported by standard deviation of the traits with grammar being the highest (SD=6) followed by reading (SD=4) followed by vocabulary (SD=3). It means that the students differed more in grammar than other traits. But when we excluded the traits, the correlational indices changed as shown in the table. In terms of descriptive statistics and correational analyses, grammar is said to be the best predictor of success. But in the light of possible evidence from consequential research and practical issues, the major role should be allocated to the reading trait as opposed to other traits. The last column in Table 5 also puts a stamp of approval on this.

Conclusions

Factor analysis revealed that the test is valid. The reading section does have clear factor structure and the items were loaded under the postulated factors except for item 91 which was loaded under the wrong factor. It was seen that for the vocabulary section, the best way of dealing with validity issues was via multitrait-multimethod design. Other methods except for inter sub-test correlations were not appropriate. The application of this procedure was a success and it showed that the test was valid. Specifically, it was seen that the same traits tested through the same methods had a higher correlation than that of the same traits having tested through different methods.

One good thing about MTMM design is that it covered all sub-skills on the test. One might say that the reading section was not involved. It might seem a truism. But vocabulary items taken from the reading comprehension passages provide a piece of evidence on the contrary. To sum up, using MTMM design was very helpful in collecting evidence for the validity of the test. It was seen that the test was valid in the light of evidence as brought by the inter sub-test correlation. The UTEPT is an example of a test that has different sub-sections which make it possible for the test to be analyzed from different perspectives.

It was seen that for the vocabulary section, the best way of dealing with validity issues was via multitrait-multimethod design. Other methods except for inter sub-test correlations were not appropriate. The application of this procedure was a success and it showed that the test was valid. Specifically, it was seen that the

same traits tested through the same methods had a higher correlation than that of the same traits having tested through different methods.

One good thing about MTMM design is that it covered all sub-skills on the test. One might say that the reading section was not involved. It might seem a truism. But vocabulary items taken from the reading comprehension passages provide a piece of evidence on the contrary. To sum up, using MTMM design was very helpful in collecting evidence for the validity of the test. It was seen that the test was valid in the light of evidence as brought by the inter sub-test correlation. The UTEPT is an example of a test that has different sub-sections which make it possible for the test to be analyzed from different perspectives.

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* NY: CUP.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8, 41-66.*

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: OUP.

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, OUP.

Baker, D. (1989). *Language testing: A critical survey and practical guide.* London: Edward Arnold.

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

Camps, J. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. *IJAL*, *13*(2), 201-221.

Cohen, A. (1984). On taking language tests: What the students report. *Language Testing*, 1, 70-81.

Hatch, E., & Lazaraton, A. (1991). *A research manual: Design and statistics for applied linguistics.* NY: Newbury House Publishers.

Kline, P. (1994). *An easy guide to factor analysis.* NY: Routledge.

Kormos, J. (1998). The use of verbal reports in L2 research. *TESOL Quarterly*, *32*, 353-363.

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

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

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. Proceedings of a colloquium at TESOL'79, Boston.*

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's factor analysis machine. *Understaning Statistics, 2*, 13-43.

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