Journal of English Language
Teaching and Learning
Year53 No. 222

On the Construct Validity of the Reading Section of the University of Tehran English Proficiency Test

Dr.Mohammad Salehi
Assistant Professor, Sharif University of Technology

Abstract

University of Tehran administers a test known as The University of Tehran English Proficiency Test (the UTEPT) to PhD candidates on a yearly basis. By definition, the test can be considered a high-stakes one. The validity of high stakes tests needs to be known (Roever, 2001). As Mesick (1988) maintains, if the validity of high stakes tests are not known, it might have some undesirable consequences for the society at large. To investigate the construct validity of the test, factor analysis was employed (Farahdy, 1983). The factor analysis of choice was exploratory factor analysis with varimx rotation strategy. In the reading section, the factor analysis produced eleven factors. Furthermore, the use of Principal Components Analysis (PCA) was complemented with Principal Axis Factoring (PAF). Overfactoring was an issue. This overfactoring can be attributed to the heterogeneous nature of items; different paradigms from which items were taken: iBT TOEFL, ILTES, and FCE.

Key Words: Construct Validity, Factor Analysis, Varimax Rotation, Principal Components Analysis, Principal Axis Factoring.

- تاریخ وصول :۸۹/۴/۱۵ تأیید نهایی:۸۹/۱۱/۲۳

1.1. Introduction

University of Tehran administers a test to PhD candidates on a yearly basis. The test is a high stakes one; almost 10,000 candidates take it. Admission tests for universities or other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations (Roever, 2001).

So far to the knowledge of the researcher, no in-depth study has ever been conducted regarding the validity of the University of Tehran English Proficiency Test (the UTEPT). According to Messick (1988), if a test has life changing implications for the individuals involved, its validity needs to be revealed. The UTEPT determines the acceptance or non-acceptance of PhD candidates into programs of instruction and attempts should be made to investigate its validity. The current study employs exploratory factor analysis to find out whether the test possesses the qualities expected of a test of this caliber.

The use of principal components analysis (PCA) is usually fraught with controversy. For

example, Farhady (1983) entertains the view the first component usurps the most of variance. Hatch and Lazaraton (1991) recommend using PCA prior to common factor analysis. In this study, PCA and common factor analysis are deemed to complement one

another.

1.2. Rationale for the Study

As it was mentioned before the UTEPT is a high stakes test. The academic lives of individuals hinge on outcomes of the test. Considering the fact that the test has been around for the past 15 years or so, and no in-depth study has been ever conducted, the current study is justified. The UTEPT consists of three distinct sections: grammar, vocabulary and reading comprehension. Out of these sub-skills, reading comprehension constitutes good grounds for investigation. It is a good predictor of the candidates' future academic careers as their future textbooks are in English which basically revolves around reading comprehension abilities. Furthermore, reading comprehension is susceptible to the influence of many a construct irrelevant factors (James Purpura, personal communication). It should come as no surprise that reading comprehension sections of measures lend themselves to differential item functioning investigations because the roles of these irrelevant factors is more palpable in this section than other sections. Field of study is a case in point. It is a truism that students majoring in different fields may perform differently in different topics of reading comprehension.

2. Review of the Related Literature

2.1. Definitions of Construct Validity

Palmer and Groot (1981) view construct validation as a theory testing procedure and distinguish it from all types of validity in which reference to a criterion is important. In their definition, the importance of exploratory factor analysis and confirmatory factor analysis is emphasized. They maintain that:

In construct validation, one validates a test not against a criterion or another test, but against a theory. To investigate construct validity, one develops or adopts a theory which one uses as a provisional explanation of test scores until, during the procedure, the theory is either supported or falsified by the results of testing the hypotheses derived from it. (p. 4)

Hughes (1989) also defines construct validity in a manner which has often been quoted by other researchers (e.g., McDonough, 1995).

A test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure…..One might hypothesize, for example, that the ability to read involves a number of sub-abilities, such as the ability to guess the meaning of unknown words from the context in which they are met. (p. 26)

An interesting point about this definition is that it can be applied to language testing per se. What Hughes implies is that reading is a multi-faceted phenomenon. There are various sub-abilities involved in the reading process. Inferencing, vocabulary, and topic identification are some of these.

2.2. Approaches to Construct Validation

There have been several approaches to test validation. A sketch of the approach of Alderson, Clapham, and Wall (1995) is the most appealing. The first approach that they talk about 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 that 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 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 including the sub-part 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 into account

test bias and actually assessing the role of background knowledge, gender, race, etc. Two of the aforementioned approaches will be dealt with here. For further information on other approaches the interested reader may refer to other sources (e.g., Zumbo, 1999).

2.2.1. Factor Analysis

Factor is a data reduction technique which reduces variables and categorizes them into components or factors. There are a good number of definitions available. Baker's (1989) definition is very concise who 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 types of factor analysis techniques: exploratory and confirmatory. In the exploratory factor analysis, we do not determine which variables should go with which factors. And in the confirmatory factor analysis we have a hunch which variables are related to which factors. As for exploratory factor analysis, Bachman (1990) maintains, "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).

2.2.2. Multitrait Multimethod (MTMM)

Perhaps the pioneers for MTMM designs can be Campbell 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 for the convergent validity, it can be maintained 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 exercised.

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.

2.2.3. The Role of Background Knowledge

Alderson et al (1985) consider the role of background knowledge as another approach by which construct validation studies can be pursued. The areas of background knowledge include but is not limited to gender, field of study and age of participants.

Zumbo (1999), among others, is of the belief that a construct validation study needs to take into account construct irrelevant factors. Detecting differential item functioning (DIF) is one way of addressing that concern.

Zumbo (1999) maintains that "DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure" (p. 12). In other words, if the total population is matched in terms of gender, let's say, and the overall performance of the two groups is almost the same as determined by the mean comparisons of the two, any differential performance on

specific items can attributed to the role of gender which is not related to the construct under study.

2.3. Research Question

The research question addressed in the current study is as follows:

Do the test items in the 'Reading Comprehension' section of the UTEPT distinctly measure various sub-skills?

3. Methodology

3.1. Participants

The participants included in the present study were 3,398 testees chosen from the total population of 8,696 testees who took the UTEPT in February 2007 (Esfand of 1385). Outliers were discarded. The participants majored in different fields of study, including physics, chemistry, theology, etc. Almost forty different majors were represented in the current study. Unfortunately, certain pieces of demographic information were not obtainable. These include but are not limited to the age of the participants. As for the number of participants that should be present in factor analysis studies, different scholars hold different views For example, Kline (1994) suggests that the number of subjects should be two times as many as the number of variables. Henson and Roberts (2006) maintain that, "It is not uncommon to find rules of thumb in the factor analytic literature; it is less common, though, to find consistency in recommendations" (p. 402). They further refer to other scholars' recommendations. For example, they mention Stevens (1996) as suggesting that "the number of participants per variable is a more appropriate way to determine sample size (ranging from 5 to 20 participants per variable). Fewer participants are needed when component saturation is high" (p. 402). Hopefully, the number of participants was large enough for factor analysis to be conducted.

3.2. Instrumentation

3.2.1. 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 (multiple choice paraphrases) has 10 items (items 46 to 55) and part two (multiple choice completion) has 10 items (items 56 to 65). The last section is concerned with reading comprehension. This section has 35 items consisting of six passages. Table 2 summarizes the three sections and parts of the UTEPT. It should be noted that the format varies from one year to the other.

Table 2. Three Different Sections of the UTEPT

Section	Method of testing	Number of items	Item Number
grammar	Multiple Choice Completion	20	1-20
	Error Identification	15	21-35
	Contextualized	5	36-40
vocabulary	Multiple Choice Paraphrases	10	46-55
	Multiple Choice Completion	10	56-65
	Contextualized	5	41-45
reading comprehension	Six short passages	35	66-100

3.3. Data Collection

It should be noted that the researcher did not have any role in the data collection process. It was obtained from the information center of the University of Tehran via a great deal of paper work. The UTEPT is constructed and administered by faculty members of the English department at the University of Tehran.

3.4. Data Analysis

3.3.1. Exploratory Factor Analysis

To answer the research question of the study, .i.e., "Do the test items in reading comprehension section of the UTEPT distinctly measure various sub-skills?" exploratory factor analysis was performed. This statistical procedure was used to extract factors in the reading comprehension section. The extraction method was Principal Components Analysis (PCA). The justifications are as follows:

1- It is mathematically simple (Kline, 1994). In other words, the algebra and computation of Principal Components Analysis is not complex.

2- The computational methods used make absolutely clear the basis of the assertions that factors account for variance and explain correlations (Kline, 1994).

3.3.3.1. Principal Axis Factoring Following Principal Components Analysis

Principal Axis Factoring followed Principal Components Analysis. The addition of this is also consonant with the call of the scholars in the field of language testing research to carry out other methods of factor analysis following or in place of Principal Components Analysis (e.g., Hatch & Lazaraton, 1991).

4. Results

4.1. A Factor Analysis Performed on the Reading Comprehension Section Using Principal Components Analysis

To answer the research question, the researcher operated on the assumption that reading is a trait which consists of sub-abilities (Hughes, 1989) and expected that factor analysis would yield some sub-abilities.

Using exploratory factor analysis (Principal Components Analysis), 11 factors were extracted. Correlations below .30 were suppressed. What follows is a list of the factors and their interpretations. Table 3 shows the extracted factors and their loadings.

It should be noted that items are revealed here because the test content varies on a yearly basis. Otherwise, it would have been unethical to expose the items to the public.

Table 3. 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				-			
q070			-		-			.361			
q071			.395		.339			.544			
Q072	.456										
Q073											
Q074							.365	.606			
Q075								.432		.504	
Q076						-					
Q077						.614				.592	.449
Q078											
Q079					.388	.694					-
Q080						.366				-	.449
Q081		.316								.511	
Q082									-		
Q083	.391								712.		
Q084									-		
Q085					.514				.334		
Q086									.464		
Q087			.327								
Q088	.552	.562									.692
Q089		.599									
Q090	.624										
Q091											
Q092							.627				
Q093					.585						
Q094	.495	.322									
Q095			.462				.322				
Q096			.386				.388				
Q097			.599								
Q098					.339		-				
Q099			-				.493				
Q100		.440	.434					-			
								.307			

4.1.1. 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 B. improved C. popular D. possible

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

4.1.2. 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 B. improved C. popular 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 has commonalities with factor one.

4.1.3. 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 and will 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 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. The items will be briefly discussed in terms of factor loadings and attempts will be made to factor name them.

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

C. where land development exceeded population growth

B. that is used as a model for land development

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 and foremost, items 99 and 71 have loaded on this factor. This is surprising because they are vocabulary items and our expectation was that they would be

loaded on the first extracted factor. The second point pertains to 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.

4.1.4. Factor Four

This factor consist of items 66, 67, 68, and 69. Factor loadings are relatively high .589, .582, .388, and .451, respectively. The items are shown:

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 easy ones. 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.

4. 1.5. 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 loads more on factor 11 than it does on this particular factor. Item 85 has a relatively high factor loading and is factor pure. By 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 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.

4. 1.6. Factor Six

Items 76, 78, and 79 came to be loaded on this factor. Item 76 is factor pure with a 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 items 76 and 78 should help us in factor naming. First, items should be closely inspected:

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 is a reference item. Probably, all items are concerned with word paraphrases.

4.1.7. 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 .years C. lots D. developers

It should come 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 bears any resemblance to this one. This factor can rightfully named a grammatical reference item.

4. 1.8. Factor Eight

Items 69, 70, 73, 74, and 99 came to be included under this factor. Item 69 is not factor pure and 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 can observe what has happened to item 70. This item has a large, albeit not the largest, factor loading. The factor is probably expecting a great contribution from the item. Next, there is item 73 with the largest factor loading of all the items and is expected to make a good contribution to the extracted factor. The last two items are not factor pure which means that they are not expected to be of any help in naming the factor. The two most contributing items, i.e., items 70 and 73 can be found 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 mot 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 safely called "understanding through paraphrase".

4. 1.9. Factor Nine

Items 82, 83, and 84 loaded on this factor. Item 82 is factor pure. Item 83 is not and also loads on factor one. So, this item is most probably a vocabulary factor. Finally, item 84 is also factor pure and accountable for explaining the most variance. Items 82 and 84 will be scrutinized to 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 behaved in the way we expected. 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".

4. 1.10. Factor 10

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 be of any help 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. Please notice 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".

4. 1.11. Factor 11

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.

4. 2. Correlation among the Factors

To gain more insights into the factors and their interpretations, factors were intercorrelated. Low correlations assume independence among factors (Preacher & MacCaluum, 2003). The results are demonstrated in Table 4.

Table 4. Correlations among the Factors

	1	2	3	4	5	6	7	8	9	10	11
1	1										
2	.049	1									
3	-.029	-.049	1								
4	.182	.017	.008	1							
5	-.108	-.067	-.079	-.060	1						
6	.114	.081	.019	.089	-.057	1					
7	-.009	-.066	.060	.071	-.074	-.008	1				
8	.139	.017	.002	.179	-.002	.113	.046	1			
9	.075	.101	-.050	.008	-.074	.066	-.047	.021	1		
10	-.029	-.001	-.031	-.030	.052	-.068	-.024	-.015	-.033	1	
11	.082	.015	-.020	.024	-.040	.038	-.022	.048	.074	-.020	1

Generally speaking, correlations are low. This is not surprising. After all, factors assume independence. Considering the rotation strategy used in the present study, i.e., varimax, it is assumed that factors should be uncorrelated which is not necessarily the case. For instance, factor four and eight are correlated.

To conclude this section, one can say the answer to the research question raised is positive on the grounds that 11 distinct factors emerged through a factor analysis. Negative correlations speak to the divergence of traits in the factors under question.

4.3. Factor Analysis on Reading Comprehension Items Using Principal Axis Factoring

Based on research in the past and criticism leveled at Principal Components Analysis (e.g., Farhady, 1983), Principal Axis Factoring was also run. The opinions about the superiority of other methods of factor analysis over principal components analysis are not entirely consistent. For example, Kline (1994) relying on Harman (1976) reminds us that with large data sets the distinction between principal components analysis and other methods of data condensation becomes virtually insignificant. Table 5 shows Pattern Matrix for Principal Axis Factoring. A comparison and contrast of these two ways of extraction methods are in order.

Table 5. Principal Axis Factoring on Reading Comprehension Items

	Component										
	1	2	3	4	5	6	7	8	9	10	11
q066				.390							
q067				.427							
q068											
q069											
q070										.360	
q071											
q072											
q073										.321	
q074											
q075											
q076					.322						
q077											
q078					-						
q079					.593						
q080									.362		
q081											
q082							.575				
q083											
q084											
q085										.	
q086											
q087											
q088			.447								
q089											
q090	.502										
q091											
q092											
q093											
q094	.381										
q095		.581									
q096											
q097											
q098						.545					
q099								.415			
q100											

4.3. 1. Factor One

PCA: Five items loaded on this factor. These items are 72, 83, 86, 90 and 94.

PAF: Two items are loaded on this factor; items 90 and 94.

Elaboration: While this factor is clearly a vocabulary factor as claimed by PCA, PAF only has considered two items as vocabulary items.

4. 3.2. Factor Two

PCA: Items 81, 88, 89, 94 and 100 loaded on this factor.

PAF: Only one item, 95, loaded on this factor.

Elaboration: Item 95 does not appear among items in PCA.

4.3.3. Factor Three

PCA**:** Items 71, 87, 95, 96, 97, and 99 loaded on this factor.

PFA: Only item 88 loaded on this factor.

Elaboration: Again the two extraction methods did not necessarily yield the same results.

4. 3.4. Factor Four

PCA: Items 66, 67, 68, and 69 loaded on this factor.

PAF: Items 66 and 67 loaded on this factor.

Elaboration: PCA and PAF seem to be in partial agreement. Both have brought these two items under the same factor. In fact, the factor loadings of items 66 and 67 in PCA are greater than they are for items 68 and 69. In a nutshell, the two have almost converged on the same conclusion.

4. 3.5. Factor Five

PCA: Items 71, 80, 85, 93 and 98 loaded on this factor.

PAF: Two items were brought under this factor; 76 and 78. None of these items can be found among the PCA items.

Elaboration: The two methods did not reveal the same results.

4. 3.6. Factor Six

PCA: Items 76, 78 and 79 loaded on this factor.

PAF: Item 98 loaded on this factor.

Elaboration: The two extraction methods did not agree at all

4.3.7. Factor Seven

PCA: Items 74, 93, 96, 97 and 99 loaded on this factor.

PAF: Item 82 was brought under this factor.

Elaboration: Again the two extraction methods have acted inconsistently.

4. 3.8. Factor Eight

PCA: Items 69, 70, 73, 74 and 99 loaded on this factor.

PAF: The only item that loaded on this factor was 99.

Elaboration: The two extraction procedures seem to be in agreement.

4. 3.9. Factor Nine

PCA: Items 82, 83, and 84 loaded on this factor.

PAF: Item 80 loaded on this factor.

Elaboration: The two methods did not achieve the same results.

4. 3.10. Factor 10

PCA: Items 75, 77, and 80 loaded on this factor.

PAF: Items 70 and 73 loaded on this factor.

Elaboration: The two procedures did not point to the same results.

4. 3.11. Factor 11

PCA: Items 79, 81, and 91 loaded on this factor or component.

PAF: No specific item loaded on this factor.

4. 4. Conclusions about the Comparison and Contrast between Two Methods of Extraction

Both extraction procedures extracted 11 factors and components. It has to be mentioned that Principal Axis Factoring did extract 11 factors, although apparently no item can be observed under factor 11. This is because of suppression level. Had the researcher accepted factor loadings lower than .30, he would have had items to come under this factor as well. A comparison can be run among the extracted factors in the two extraction procedures. Under factor one in PCA there are four items: 72, 86, 90, and 94. In PAF, there are two items, 90 and 94 which were loaded under this factor. In other words, there is no difference in terms of the underlying traits as illustrated by the two factor analysis methods, because both extracted vocabulary items. For factor two in PAF, there was item 95. This item was not factor pure in PCA. The item may be said to tap implicit knowledge in the text. It is shown below:

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

The reader may recall that it was named an inference item in PCA. One can conclude that the two extraction methods treated this item in the same way.

In PAF, item 88 as shown below came to be included under factor three. The reader may recall that the same item along with items 89 and 100 loaded on one separate factor.

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

The reader may further recall that the items were referred to as main idea items. The same label applies to this particular item. As for factor four, the two extraction methods were somewhat consistent. Items 66, 67, 68, and 69 came to be included under this factor in PCA. PAF brought items 66 and 67 under this factor with the factor loadings of .30 and above. Items 68 and 69 were not included under this factor with the factor loading of .30 and above. All in all, it can be concluded that the two methods did differ to some extent. It is a point where one has to disagree with researchers who claim that with large data sets of which the current study is an example, the difference in the choice of one extraction method over another becomes insignificant (Kline, 1994). Factor six consists of an item that is special. This is technically called "drag and insert" item. PAF has treated it separately. In PCA, this item is not factor pure. As for items 82 and 84, they were factor pure and loaded on the same factor in PCA. PAF did bring item 82 under one separate factor. But item 84 was not loaded.

Surprisingly, item 99 which is a vocabulary item was loaded on factor eight. We say surprisingly, because it is a vocabulary item and was not expected to come under this factor. It is to be recalled that vocabulary items belonged to factor one in both extraction methods. Item 80, along with items 75 and 77, belonged to the same factor in PCA, while item 80 went under the factor alone. As for factor 10, the two extraction methods funtioned identically. They brought items 70

and 73 under the same factor. The last factor is not going to be dwelt on because of low factor loading.

4. Discussion

There is the problem of over factoring because 35 items lend themselves to 11 factors. This is, however, justifiable on the grounds that the reading passages were extremely heterogeneous by nature, meaning that the test constructor opted for an amalgamation of different orientations in language testing. It should be noted that this pertains to methods not traits, but methods and traits are sometimes indistinguishable (Stevenson, 1981). In other words if we can operationally define orientation (FCE, IELTS,

TOEFL) as methods then it is entirely possible that methods or orientations might have

induced error into the process. According to Bachman (1990), the performance of test takers on a test might be more of an engagement with methods than with traits. The factor analysis might be appropriate but the distinction between traits and methods may become fuzzy.

Another obvious problem is under factorability in the sense that some factors were

not represented. According to Messick (1989), this is referred to as construct under representation. To exemplify it, topic prediction was never represented in the reading comprehension questions. The UTEPT is an example of a proficiency test not an achievement one. The point is that a proficiency test should embody the constructs of a test, but in this study some constructs were under represented. According to Weir (1990) this is referred to as an apriori validation whereby the test maker operates on preconceived assumptions. In other words, he is equipped with a table of specification.

Weir also talks about posteriori approach to validation of which this study is an example.

While the researcher could control the latter he could not exercise any control over the former. With no table of specifications available to the researcher, it was very difficult to

see what the sub-skills were intended by the test constructor.

Another point should be made about the factor analysis. Carroll (1983) reminds us that each factor extracted should be represented by at least three variables. In this study, it

could not be materialized. As it can be observed, only four out of 11 factors do meet the criterion as set by Carroll (1983). Of course some factors could have had more than the current number of variables if the researcher had accepted lower factor loadings. This is in

close alignment with overfactoring. It may be the case that the items are so different in terms of the underlying traits.

Correlations among factors provided evidence for the distinctness of traits or factors. Generally speaking, the intercorrelational indices were low bearing testimony to the

separateness of traits. In language testing, zero correlations are not a possibility as some relationship between variables is to be expected no matter how distinct the factors are. Relatively high correlations emerged between factor 1 and 4 on the one hand, and 4 and 8 on he other. The reader recalls that factor 1 is a vocabulary factor and factor 4 is a factor which was named directly answerable question factor. It comes as no surprise that these two factors share commonalities. To be more specific, factor 4 entails a vocabulary item. On the other hand, a relatively high correlational index that emerged between factor 4 and 8 is not explainable because the two factors tap almost unrelated concepts.

5. Conclusions

It can be concluded that factor analysis proved a robust tool in the investigation of construct validity. Principal Components Analysis (PCA) could easily delineate factors in the section. Eleven factors were extracted out of 35 reading comprehension items. Considering the fact that it was bizarre for this number of factors to be extracted, it was deemed appropriate to employ another method of extraction. Therefore, Principal Axis Factoring was performed. Some scholars believe that due to shortcomings associated with PCA, other methods of factor analysis should be used instead of or following PCA (Hatch & Lazaraton, 1991).

The results of the study should be treated with caution. First and foremost, there was the issue of overfactoring. It can be justified on the grounds that the test developer used

different orientations in language testing. In other words, passages from ILTES, TOEFL, and FCE were included. The passages are a melting pot of various approaches to language

testing. Factor analysis is supposed to delineate underlying traits not the methods employed. Apparently the same traits

were tested using different methods. The study calls for a research in terms of comparability of TOEFL and ILTES (see also

Chalhoube-Deville & Turner, 2000). In terms of performance of testees, they diverged when it came to the two orientations. It remains to be seen whether the two orientations would yield the same results. In this study, they did not. Second of all, with truncated subjects, some valuable information may have been lost

6. Suggestions for Further Research

1- Differential item functioning (DIF) is a ripe area for research (Geranpayeh & Kunnan, 2007). What it claims is that individuals with the same ability level in terms of overall score may perform differently on certain items. This difference can be triggered by their fields of study, for instance.

2- Another area that can be explored is a criterion-related analysis study. The UTEPT can be administered along with a criterion measure to the same group of testees. The criterion measure could be a version of CBT which taps similar constructs as the UTEPT does. A high and positive Pearson Product Moment Correlation can speak to the validity of the UTEPT.

3- Test taking strategies is another promising ground for research. The current study did use test taking strategy research. But any potential researcher can use questionnaires to elicit the test taking behavior of testees. The questionnaires can be administered right after testees take tests. It can be seen whether different strategies are used for different item types in the reading comprehension sections. Both anecdotal evidence and research findings support the use of test taking strategies with the reading section of a proficiency measure (e.g., Cohen & Upton, 2007).

4- Additionally, within the framework of test taking strategy research, the role of proficiency can be gauged.

5- A potential research project can be geared towards the notion of construct validity as elaborated by Bachman and Palmer (1996). To the best of the knowledge of the researcher the area has not been explored by Iranian student researchers.

6- The researcher used factor analysis to provide evidence for the validity of a high stake test. Other research tools, like IRT models, might be applied to collect evidence for the validity of the test.

7- An interesting research study could be to factor analyze the current test and the criterion measure (see number 2 above).

The factors extracted from the two tests could be correlated with one another to gain insights into the underlying constructs of the two tests (Kline, 1994).

7. Implications of the Study

It is not a bad idea to put a variety of items on a test. This, according to Henning (1987), would enhance content validity at the cost of jeopardizing internal consistency. But, an investigation of the items showed that the reading comprehension items came from different paradigms like TOEFL, FCE and even IELTS. This can create problems as far as test fairness is concerned. A testee who has helped himself to a little bit of everything (TOEFL, FCE, etc) might unfairly outperform a student who has devoted himself to one of these, but intensively so. An analysis of the items showed that it is best not to include heterogeneous items in a test battery like the one under the investigation.

Furthermore, factor analysis is often frowned upon as an old statistical device (Stanely Muleik, personal communication). But the results of the current study showed that it can be optimally employed in the case of a high stakes test. Last but not the least, it is best to complement the use of PCA with common factor analysis.

8. **(De) limitations of the Study**

Here are some of the (de)limitations of the current study:

1- Although gender can play a role in the choice of test taking strategies, this study does not take this into account.

2- Another variable that is important and could have been investigated is the role of the field of study. This is a quite promising area for research. But it was ignored because of time- related constraints.

3- Confirmatory factor analysis could have also revealed insights into the factor structure of a test.

4- Consequential validity is the concern of Messick (1989). The researcher could have conducted interviews to elicit ideas from testees to see if they see the contents of what they sat for as an entrance test have any correspondence to what they are exposed to in their courses.

5- This is a post-hoc analysis of a high stake test. A better analysis would have been an a priori analysis.

References

Alderson, C.(2000). *Assessing reading.* Cambridge: Cambridge University Press.

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

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & Braun, H. (Eds.). *Test validity* (pp. 19-32). Hillsdale, NJ: Erbaum.

Bachman, L. (1990b). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

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

Brown, J. D. (1988). *Understanding research in second language learning*. New York: Cambridge University Press.

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

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

Carroll, J. B.(1983). Psychometric theory and language testing. In W. J. Oller, (Ed.), *Issues in langauge testing research* (pp. 80-107). Rowely, MA: Newbury House.

Chalhoube-Deville, A., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS and TOEFL. *System, 28*(4), 523-539.

Cohen, A., & Upton, T. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph No. 33). Princeton, NJ: Educational Testing Service.

Farhady, H.(1983). On the plausibility of the unitary langauge proficiency factor. In

W. J. Oller, (Ed.), *Issues in langauge testing research* (pp. 11-28). Rowely, MA: Newbury House.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced

English examination. *Language Assessment Quarterly, 4* (2), 190-22.

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

Henning, G. (1987). *A guide to language testing.* Cambridge: Newbury House Publishers.

Henson, R. K., & Roberts, J. K. (2006) Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393-416.

Hughes, A. (1989). *Testing for language teachers.* New York: Cambridge University Press.

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

McDonough, H. (1995). *Strategy and skill in learning a foreign language.* London: Edward Arnold.

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.

Messick, S.(1989). Validity. In R.L. Linn (Ed.), in *Educational Measurement* (3rd. edition, pp. 13-103). New York: American Council on Education and Macmillan.

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

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.

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

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

Weir, J. C. (1990). *Communicative language testing.* Hempstead: Prentice Hall

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.