On the Factor Structure of a Reading Comprehension Test

Mohammad Salehi (PhD)
Assistant Professor, Sharif University of Technology
E-mail: m_salehi@sharif.ir

Received: January 26, 2010 Accepted: February 20, 2010 doi:10.5539/elt.v4n2p242

Abstract

To investigate the construct validly of a section of a high stakes test, an exploratory factor analysis using principal components analysis was employed. The rotation used was varimax with the suppression level of .30. Eleven factors were extracted out of 35 reading comprehension items. The fact that these factors emerged speak to the construct validity of the test. However the problem of over-factoring was obvious. This may be attributable to different paradigms of testing on which the items were based. In other words, the test constructer opted for passages from TOEFL, FCE and IELTS books with much alteration.

Keywords: Construct validity, Exploratory factor analysis, Varimax rotation, High stakes tests, Principal components analysis

1. Introduction

University of Tehran administers a proficiency test to PhD candidates on a yearly basis. The test can be considered a high stakes one by the virtue of the fact that 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). According to Messick (1988), if the validity of a test is not known, it might have undesirable consequences for the society at large.

The purpose of the current study is to investigate the factor structure of the reading section of University of Tehran English Proficiency Test (the UTEPT). While the UTEPT plays a key role in the academic lives of individuals, no in-depth study has ever been conducted regarding the validity of the test. The only study is that of Zand Karimi (2005). The study, however, has methodological flaws, not the least of which is the inappropriate use of Principal Components Analysis (PCA). Specifically, PCA has been applied with no reference to loading patterns. The current study attempts to shed more light on the factor structure of the reading comprehension of the test. Information on other sections can be found elsewhere (Rezaee and Salehi, 2008, and Salehi and Rezaee, 2008).

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 underscored. 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's (1989) definition 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 measur. 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 being some of them.

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 verify 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 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 yield a positive correlation with the total test. Still, another problem may arise; the correlation of any sub-part with the total test with 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 (MTMM) 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.

Out of the approaches mentioned above, three needs more elaboration that was already given. One is factor analysis. The other one is multitrait multimethod. And finally he last one is the role of background knowledge to assess the role of bias in the testing process.

2.2.1 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 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

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

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

As 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.

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

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

Zumbo (ibid) 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)

2.3 The Research Question

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

Do the test items in the ‘Reading Comprehension’ sections of the UTEPT distinctly measure various sub-skills?

3. Methodology

3.1 Participants

The participants in the present study were 3,398 testees chosen 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. 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).

Usually, "the larger the better" sounds all too familiar. But the rule of thumb is two times as many as the number of variables. In the current study, there were 100 variables corresponding to the number of test items. However, the sample size in the current study exceeded the criterion level. There were 3,398 participants. The sample size can play a very crucial role. It is even said that with larger samples, the distinction between using various types of factor analytic techniques becomes insignificant (see, Kline, 1994).

In SPSS, there is a convenient option offered to check whether the sample is big enough. This is the Kaiser-Meyer-Olkin measure of sampling adequacy. The sample is adequate if the value of KMO is greater than 0.5. To check the adequacy of sampling, KMO was performed and the results showed that the KMO test of adequacy of sampling was .753 for the total test. It is greater than .5. So the test is adequate in terms of sampling.

Variance is a crucial factor in test validation which primarily affects reliability. In other words, the higher the variance, the more the reliability. There is a substantial body of evidence for the importance of reliability as a prerequisite for validity. Kline (1994) maintains that heterogeneity is a crucial issue in factor analysis. This assumption was met because the variance of this test was 39.873.

Taking into account the criteria in the literature, it becomes obvious that the current study meets the benchmarks as set by factor analysts in the social sciences.

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 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.

3.3 Data Analysis

To answer the research question of the study, .i.e., "Do the test items in the reading comprehension sections of the UTEPT distinctly measure various sub-skills?" an 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).

4. Results and Discussion

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.

Having used 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. All the 11 components accounted for 41.96 percent of the total variance. Table 2 shows the extracted factors and their loadings. 4.1.1. Factor One

Items 72, 83, 86, 90 and 94 loaded on factor one which is 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. 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.

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. 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.

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.

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.

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.

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 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 can be scrutinized to see if our prediction about the characteristic of item 83 is borne out.

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.

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 has surprisingly loaded on this factor. It is the point where factor analysis should be combined with logic.

5. 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. As Bachman (1990) puts it, the performance of test takers on the test might be more of 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 among the factors. According to Messick, 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 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. To strip it into simple language, 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) is of the opinion 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. .

6. Conclusions

It can be concluded that factor analysis proved as 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. The results of the study should be treated with circumspection. 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. In terms of performance of testees, they diverged when it came to the two orientations. It remains to be seen whether the two orientations yield the same results. In this study, they did not. Second of all, truncated subjects coud have altered the results.

References

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

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

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

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

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: CUP

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.

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.

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

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

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. *Understaning Statistics*, 2, 13-43.

Rezaee, A. A. and Salehi, M. (2009). The construct validity of the University of Tehran

English Proficiency Test: A Multitrait Multimethod Approach. *Journal of Teaching English Language and Literature Society of Iran, 2*(8), 93-110.

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

Salehi, M. and Rezaee, A. A. (2009). On the factor structure of the grammar section of the University of Tehran English Proficiency Test. *Indian Journal of Applied Linguistics, 35*(2), 169-187.

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

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

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)

Table 2. 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	.449
Q078					.388	.694					
Q079						.366					-.449
Q080										-.511	
Q081		.316									
Q082									-712.		
Q083	.391								-.334		
Q084					.514				.464		
Q085											
Q086			.327								
Q087											.692
Q088	.552	.562									
Q089		.599									
Q090	.624										
Q091											
Q092					.585		.627				
Q093											
Q094	.495	.322	.462								
Q095			.386				.322				
Q096			.599				.388				
Q097					.339						
Q098			-.434				-.493				
Q099								-.307			
Q100		.440									