The Factor Structure of a Written English Proficiency Test: A Structural Equation Modeling Approach

Seyyed Mohammad Alavi
University of Tehran, Iran
smalavi@ut.ac.ir

Shiva Kaivanpanah
University of Tehran, Iran
shkaivan@ut.ac.ir

Akram Nayernia
University of Tehran, Iran
nayernia@ut.ac.ir

Abstract

The present study examined the factor structure of the University of Tehran English Proficiency Test (UTEPT) that aims to examine test takers' knowledge of grammar, vocabulary, and reading comprehension. A Structural Equation Modelling (SEM) approach was used to analyse the responses of participants (N= 850) to a 2010 version of the test. A higher-order model was postulated to test if the underlying factor structure, obtained in a data-driven manner, corresponds with the proposed structure of the test. The results revealed an appropriate model fit with the data, pointing to the fact that the three sections of UTEPT, i.e., structure, vocabulary, and reading, and their sub-components, except for the restatement section of reading, are good indicators of written language proficiency as assessed by the UTEPT. It was also found that the three sections assess distinctive constructs. The findings suggest that UTEPT is a valid measure of the written language proficiency of Ph.D. applicants to University of Tehran.

Keywords: Language Proficiency, University of Tehran English Proficiency Test (UTEPT), Factor Structure, Structural Equation Modelling

Received: February 2010; Accepted: January 2011

1. Introduction

Proficiency in a second language is one of the most fundamental concepts in Applied Linguistics, and accordingly it is a subject of ongoing and intense debate. Often this debate is about competing theories or models of second language proficiency and its development (Canale & Swain, 1980; Bachman, 1990).

Providing a definition of language proficiency is challenging as any definition necessarily relies on a model, a theory, or a description of language proficiency. Canale and Swain (1980) defined language proficiency as an individual's general communicative competence in the target language environment. Bachman (1990, p. 16) defines language proficiency as "knowledge, competence, or ability in the use of a language, irrespective of how, where, or under what conditions it has been acquired." Proficiency, according to Pasternak and Bailey (2004, p. 163), "is not necessarily equated with nativeness, and certainly not all native speakers are equally skilled users of English. There are varying degrees of proficiency: being proficient is a continuum, rather than an either-or proposition". A close look at these definitions reveals that the exact nature of language proficiency or language ability has undergone some dramatic changes over the past few decades. It thus demands further investigations.

The question of whether language ability is unitary or divisible into components has been of interest to applied linguists (Sawaki, Stricker, & Ornaje, 2009). This issue has gained importance when Oller (1978) proposed the unitary trait hypothesis. Oller (1978) claimed that there exists an internalized grammar, or *expectancy grammar*, which allows for efficient, on-line processing of information and creative use of the language. He also hypothesized that language ability can be accounted for by a single trait. Strong

support for Oller's claim was obtained in Principal Component Analyses of a variety of English language tests in multiple modalities (e.g., Oller, 1978; Oller & Hinofotis, 1980). However, Oller's hypothesis was questioned by other researchers (e.g., Carroll, 1993; Farhady, 2005). Subsequent studies, in which more powerful factor analytic approaches were used, refuted the most extreme version of the unitary trait hypothesis assuming that one general factor sufficiently accounts for all of the common variances in language tests (Bachman & Palmer, 1981, 1982; Carroll, 1993; Kunnan, 1995).

Valdés and Figueroa (as cited in Vecchio & Guerrero, 1995) indicate that what it means to know a language goes beyond simplistic views of good pronunciation, "correct" grammar, and even mastery of rules of politeness. Knowing a language and knowing how to use a language involves a mastery and control of a large number of interdependent components and elements that interact with one another and that are affected by the nature of the situation in which communication takes place. Oller and Damico (1991) state that the nature and specification of language proficiency have not been determined and language education researchers continue the debate about the issues related to language proficiency.

Language testing researchers seem to agree on a multi-componential nature of language ability where a general factor exits together with some smaller factors (Oller, 1983; Carroll, 1993). Nevertheless, the exact factor structures of language proficiency is the subject of intense debate. While some studies found correlated first-order factors (e.g., Bachman & Palmer, 1981; Kunnan, 1995), others found first-order factors as well as a higher-order general factor (Bachman & Palmer, 1982; Sasaki, 1996; Shin, 2005).

In general, the assumption that language ability is a "unitary competence" (Oller, 1978) has gradually been replaced by the belief that language

competence is more complex and consists of multiple inter-correlated abilities and strategies (Bachman, 1990). One representative example of this multi-component structure is the three-level hierarchical model (Bachman & Palmer, 1996) which assumes that a proficient language speaker should not only demonstrate the structural knowledge of a target language but should also have the necessary strategies to apply that knowledge effectively in actual use (Zhang, 2010). Considering the debate on the divisibility of language proficiency into skills and components, the present study intends to use Structural Equation Modeling (SEM) approach to investigate the factor structure of UTEPT.

2. Review of the Literature

2.1. Models of Language Proficiency

Over the years, different models and theories have been proposed to account for the nature of language proficiency. Along with theoretical developments, attempts have been made to provide operational definitions of language proficiency, communicative competence, and their components. Some of these definitions have led to the development of language ability models ranging along a continuum with multidimensional models at the one end and unidimensional one at the other, and some moderate models in between (Farhady & Abbasian, 2000).

Oller (1978), inspired by Spolsky's (1973) concept of overall language proficiency, proposed that a single general language proficiency factor, referred to as "g" factor, accounts for a performance on a variety of language tests. However, the strong version of his unitary trait hypothesis has been criticized for its methodological and theoretical drawbacks. For example, Vollmer and

Sang (1983) point out that Principal Components Analysis tends to overestimate the significance of the first factor by not partitioning the total amount of test variance into common, test-specific, and error variance. Also, Farhady (1983) questioned the implication of Principal Component Analysis instead of Principal Factor Analysis. This hypothesis was also challenged by Alderson (1981) on theoretical grounds. Alderson (1981) believed that accepting one underlying proficiency factor would lead to the assumption that there was no difference among different knowledge components.

The multidimensional model involves two versions, i.e., the strong and the weak version. The strong version assumed 16 components for the total language proficiency, and the weak version speculated four skills as the dimensions of language ability (Vollmer, 1983). Other arguments have also been made about the dimension of language proficiency. For example, Cummins (1984) argues that the nature of language proficiency has been understood by some researchers as involving 64 separate language components. The multidimensional model was criticized on the grounds that it failed to accommodate for the relationship among the components and skills (Bachman, 1990). It also ignored the full context of discourse and the situation of language use (Vollmer & Sang, 1983). In addition, it has been indicated that language ability depends upon factors such as test taker characteristics, test rubrics, test method, item format, and the level of language proficiency, which are believed to be outside the scope of language itself, (Vollmer, 1983; Hughes & Porter, 1983; Alderson, 1986, 1991; Milanovic, 1988; Anivan, 1991).

Other subsequent studies investigating the nature of L2 proficiency have found that language proficiency is multi-componential. In general, there is a consensus that language proficiency consists of one higher-order factor and

several distinct first-order ability factors (Bachman & Palmer, 1981; 1982; Carroll, 1993; Bachman et al., 1995; Sasaki, 1996).

To be more specific, one needs to mention Canale and Swain's (1980) model of "communicative competence" as the first and most influential model of language proficiency. They distinguished "grammatical competence" from "sociolinguistic competence". In this model, grammatical competence consists of lexis, morphology, sentence-grammar semantics, and phonology and sociolinguistic competence includes sociocultural rules and rules of discourse.

One representative example of multi-component structure of language proficiency is the three-tier hierarchical model proposed by Bachman and Palmer (1996). According to this model, top tier consists of language knowledge and strategic competence. At the second tier, the knowledge component can be further divided into organizational knowledge and pragmatic knowledge. Meanwhile, strategic competence is composed of strategies used in goal setting, assessment, and planning. Finally, at the bottom tier, organizational knowledge can be expressed as either grammatical knowledge or textual knowledge, while pragmatic knowledge encompasses functional or sociolinguistic knowledge. Based on this model, a proficient language speaker should not only demonstrate the structural knowledge of a target language but should also have the necessary strategies to implement that knowledge effectively in actual use.

Having criticized Canale and Swain's (1980) model for the lack of a serious endeavor to generate detailed specifications of communicative language ability and that of Bachman and Palmer (1996) for relating language ability only to the context of language testing, Celce-Murcia, Dornyei, & Thurrell (1995), proposed a detailed description of communicative competence. Their model is composed of five components: discourse competence at the center of the

model, actional competence, linguistic competence, socio-cultural competence, and strategic competence. Despite all efforts to formulate language ability as consisting of various components and strategies, as Zhang (2010) argues, most practitioners in the field of language teaching and testing follow the traditional definition of proficiency whereby language proficiency comprises linguistic skills in the four core curricular areas: listening, speaking, reading, and writing.

2.2. The Structure of Language Proficiency

Bachman and Palmer (1981) investigated the construct validity of Foreign Service (FSI) oral interview through multitrait-multimethod matrix. This test, originally designed to evaluate the language proficiency of members of the US Foreign Services, evaluates not only language proficiency, but also communication and interpersonal skills. They reported strong support for the distinctness of speaking and reading as traits, and rejected the unitary trait hypothesis of language proficiency. However, their causal models indicated a sizable portion of communality in all the measures leading to a rejection of the completely divisible trait hypothesis.

To examine the construct validation of communicative proficiency, Bachman and Palmer (1982) posited three distinct traits –linguistic competence, pragmatic competence, and sociolinguistic competence– as the components of communicative competence. At the same time, they argued for a substantial general factor affecting all measures of the study. Having used confirmatory factor analysis and simultaneous multi-group covariance structure analyses, Bae and Bachman (1998) investigated the factorial distinctness of listening and reading comprehension skills and the equivalence of factor structure across two groups of language learners. They found that the two receptive skills were factorially separable while having a high correlation with

each other. They concluded that high correlation between the two skills is evidence for the same underlying factor pattern.

Concerning the divisibility of comprehension subskills measured in L2 listening and reading tests, Song (2008) investigated the factor structure of the Web-based English as a Second Language Placement Exam (WB-ESLPE) employing a SEM approach. In particular, he intended to find first, to what extent do the WB-ESLPE listening and reading items measure different comprehension sub-skills and to what extent can L2 listening and reading be considered similar or different, with regard to the divisibility of comprehension sub-skills. He found that the WB-ESLPE listening and reading items measure two or three sub-skills, and that while L2 listening and reading might share a common comprehension process, they may be distinct in the decoding processes involved due to the difference in mode of presentation. He argued that this divisibility of sub-skills in L2 comprehension tests might depend on the test takers' L2 proficiency as well as on the task characteristics of the test.

Stricker, Rock, and Lee (2005) studied the factor structure of the LanguEdge test using confirmatory factor analysis. The LanguEdge courseware (ETS, 2002) is intended to improve the learning of English as a Second Language (ESL) by providing classroom assessments of communicative skills. LanguEdge consists of two forms of a full length, computer-administered linear ESL test and supplementary materials. They found that the four sections of the LanguEdge test represented two distinct but correlated factors, Speaking, and a fusion of Listening, Reading, and Writing, not four factors corresponding to the sections of the test.

Eckes and Grotjahn (2006), reporting a study on the construct validity of C-tests, argued that language proficiency was divisible into more specific constructs. They, using Rasch measurement modeling and confirmatory factor

analysis, concluded that C-test was unidimensional instrument measuring a single dimension.

Sawaki et al. (2009) investigated the factor structure of the Test of English as a Foreign Language ™ Internet-based test (TOEFL®iBT). They identified a higher-order factor model with a higher-order general factor (ESL/EFL ability) and four first-order factors for reading, listening, speaking, and writing. Their results supports the current practice of reporting a total score and four scores corresponding to the modalities for the test, as well as the test design that permits the integrated tasks to contribute only to the scores of the test modalities.

Having employed Structural Equation Modeling (SEM) in the study of receptive skills, i.e., reading and listening comprehension and intelligence, Schroeders, Wilhelm, and Bucholtz (2010) investigated the dimensionality of language proficiency. They argued that the high overlap between foreign language comprehension measures and between crystallized intelligence and language comprehension ability can be taken as support for a unidimensional interpretation.

In'nami and Koizumi (2011) investigated the factor structure of the TOEIC in the Listening and the Reading Modules. In order to investigate the separate contribution that each test subcomponent provides to the validity of the holistic test, they devised four different models including the uncorrelated, correlated, Higher-order and the unitary model. They discovered that distinctive but correlated factors of listening and reading. Their findings support the notion of divisibility of language skills.

Considering the paucity of validation studies on the UTEPT, especially those examining its factor structure, this study investigated the factor structure of the UTEPT. It mainly attempted to investigate empirically if the underlying

factor structure obtained in data-driven manner corresponds with the proposed structure of the UTEPT. To guide this study, the following research question was developed. Does the underlying factor structure, obtained in data-driven manner, correspond with the proposed structure of the UTEPT?

3. Methodology

The research methodology used in this study was developed on the bases of the most frequent SEM methodologies in the field of language testing and educational measurement. In'nami and Koizumi (2011) used a much similar methodology to investigate the factor structure of the TOEIC in the Listening and the Reading Modules. Their methodology was recognized as a SEM methodology where models of various weights were developed for measurement purposes. In'nami and Koizumi (2011) were not interested in any structural model in which the place of the assessed components on a rather comprehensive model would be triggering issues of test validity. However, they suggested measurement models that could be used to weigh the loads of the intended factors into their associated factor or construct.

3.1. Participants

The data included the scores of 850 participants chosen from the total population of 3000 participants who took the UTEPT in October 2010. The primary sample (raw data) included around nine hundred samples (items). The data were then analyzed for missing data, positive and negative outliers and random errors of data recording operators (Tabachnick & Fidell, 2007). The data exploration phase yielded a rather normally distributed data set which was qualified for the development of SEM measurement and structural models for

further investigation of fitness in accordance with the adopted methodology (Byrne, 2010).

Sample size in SEM methodology has been a challenging issue for the past few decades. There are studies with sample sizes as large as thousands along with published research reports of investigated samples as small as a few hundred (Bentler & Yuan, 1999). The minimum sample size to deploy the SEM methodology requires up to fifteen cases, i.e., items, per parameter to be estimated (Byrne, 2010). The studies published in language testing, either using priory item banks or collecting field data, vary in the number of participants. For instance, Bae and Bachman's study included around nine hundred subjects (Bae & Bachman, 1998; Hoe, 2008).

In the current study, there are 14 fixed parameters that could be estimated for either regression weights or factor loading. Furthermore, the measures illustrated in the final combined model represent another set of 20 parameters, mainly associated error terms to these measures, which upgrade the number of the parameters to be estimated up to 34 parameters (Table 1). Therefore, this study needs a minimum number of 34X15=510 participants to satisfy the SEM sampling requirements (Hoe, 2008).

Table1. Parameter Summary

	Weights	Covariances	Variances	Means	Intercepts	Total
Fixed	14	0	0	0	0	14
Labeled	0	0	0	0	0	0
Unlabeled	6	3	11	0	0	20
Total	20	3	11	0	0	34

3.1. Instrumentation

Three equivalents forms of the UTEPT, each consists of 100 items, were used in this study. The test battery contains three sections of the test are structure, vocabulary, and reading comprehension. The structure section includes 30 items. The first 15 items are multiple choice completion items (Structure 1). The second 10 items are written expression where the test takers need to identify the erroneous part of a sentence (Structure 2). The last 5 items, i.e., grammar in context, require the participants to select an item from among alternatives to complete the text, 10 items (Structure 3). The vocabulary section includes 35 questions; for 30 items the candidates are required to choose the most appropriate equivalents/synonyms for the underlined words (Vocabulary 1) and for the other 5 items the test takers need to select the most suitable and appropriate word among choices provided to fill in the blanks (Vocabulary 2). The reading comprehension section includes 35 questions – 30 multiple choice reading comprehension questions (Reading 1) and 5 restatement items (Reading 2).

3.2. Data Analysis

Following In'nami and Koizumi (2011), to investigate empirically if the underlying factor structure obtained in data-driven manner corresponds with the proposed structure of the UTEPT, the three language skills and components were used to develop a higher-order model of construct. This model aimed to find if the underlying factor structure, obtained in data-driven manner, corresponds to the proposed structure of the UTEPT. The items were parcelled (Byrne, 2010; Tabachinic & Fidell, 2007) and used as measures of

different first-order constructs in this model. In other words, the first-order model was composed into a single model of multiple constructs contributing to a single higher-order level of constructs (Figure 1).

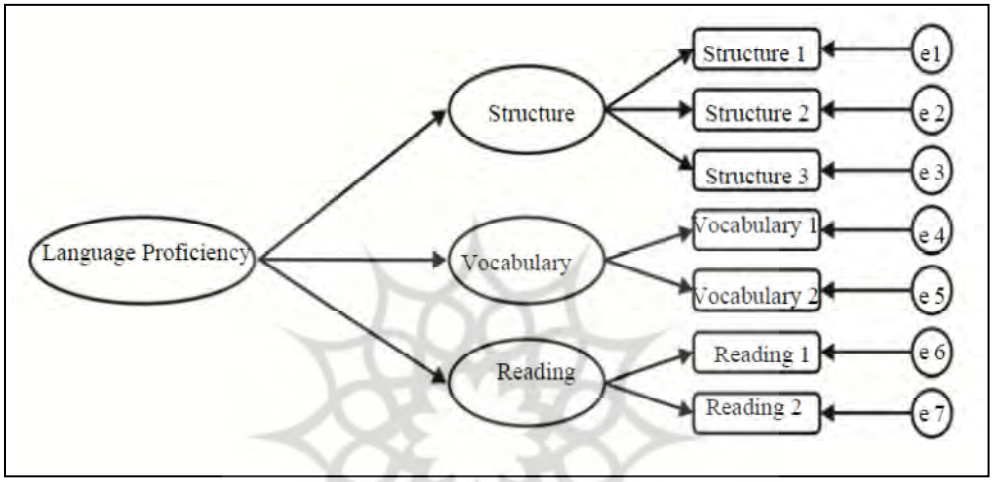

Figure 1. Higher-order Model of Language Proficiency as Measured by UTEPT

4. Results

4.1. Descriptive Statistics

Table 2 presents the descriptive statistics of the items. All the values for skewness and kurtosis were within |3.30| (*z*score at $p < .01$), which suggests univariate normality of the data (In'nami & Koizumi, 2011).

Table 2. Descriptive Statistics of Sections of UTEPT

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness
Structure 1	848	1.00	14.00	7.70	2.75	-0.04
Structure 2	848	.00	10.00	4.85	2.04	-0.09
Structure 3	848	.00	10.00	5.35	2.05	0.02
Vocabulary 1	848	.00	15.00	8.08	2.50	-0.21
Vocabulary 2	848	.00	14.00	8.03	2.65	-0.26
Reading 1	848	.00	31.00	15.21	4.63	-0.07
Reading 2	848	.00	5.00	2.01	1.19	0.32

4.2. Testing the Model

The primary model, as appeared in Figure 1, was tested against the data. The primary model was the default model in which no modifications were made and the raw data were used to test the possibility of arriving at a satisfactory model. After making necessary modifications and including more relationships in the model, the model was estimated again. As presented in Table 3, the results of model estimate for the data, indicates the higher-order model fits the data well (X^2= 2.11, df= 9, $p<0.05$). The other statistics also showed a good model fit with the data (CFI[1]= 0.995, NFI[2] = 0.99, RMSEA[3]= 0.03). CFI (Comparative Fit Index) and NFI (Normed Fit Index) are goodness of fit indices which should be above 0.90 in order for the model to be acceptable. RMSEA (Root Mean Square Residual) is badness of fit index and must be 0 and 0.1. This model was confirmed to be substantially acceptable and meaningful.

[1] Comparative Fit Index

[2] Normed Fit Index

[3] Root Mean Square Residual

Table3. Model Fit Indices

	X^2	*df*	CFI	NFI	RMSEA
Higher order model	2.11	9	0.99	0.99	0.36

As Table 4 shows, all the factor loadings, except for the construct of reading 2 with a factor loading was 0.36, ranging from 0.65 to 0.70, were statistically significant. The measurement model and the factor loadings are presented in Fig. 2.

Table 4. Factor Loadings of the Constructs in the Model

	Estimate	S.E.	C.R.	P
Structure 1	0.65	0.09	81.51	P<0.5
Structure 2	0.68	0.07	69.23	P<0.5
Structure 3	0.68	0.07	75.85	P<0.5
Vocabulary 1	0.68	0.08	93.78	P<0.5
Vocabulary 2	0.69	0.09	88.30	P<0.5
Reading 1	0.70	0.15	95.67	P<0.5
Reading 2	0.36	0.04	49.14	P<0.5

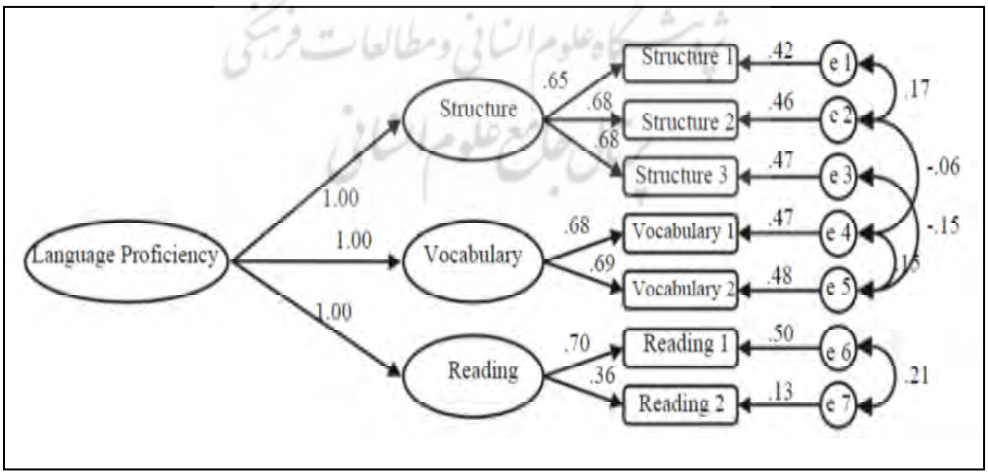

Figure 2. Factor Loadings of the Constructs in the Model

5. Discussions and Conclusions

The objective of the present study was to investigate empirically if the underlying factor structure obtained in data-driven manner corresponds to the proposed structure as suggested in the UTEPT. The model was developed from a theoretical perspective and was supposedly fitting the underlying structure of the constructs and sub-constructs of the written English language proficiency as measured by the UTEPT. There were parameters of interest in the model to be estimated through fitting the measurement models, the structural model and the final model which was tested for indices of model fitness upon the deployment of minimal model fitting measures.

Three test formats are employed in the grammar section of the UTEPT. Therefore, the module has appeared in the analysis of the present research as parcels of multiple choice grammar test items, written expression items, and grammar items in a context. The associated parcels reveal an average load of 68% into the developed construct of writing factor. Therefore, it can be justifiable to claim that the grammar section of the UTEPT can measure the function of English grammar in candidates' ability to read and understand the academic written English texts. The accumulated measures of writing ability have supported the construct of writing as measured in the UTEPT in the measurement model. The model, as appeared in Figure 2, was found satisfactorily loading into the main construct of written English language proficiency.

The measurement model of the final underlying factor structure postulated for the vocabulary in the UTEPT has been tested at two separate parcels. In the first parcel, vocabulary items in context (vocabulary 1), the candidate is required to choose one of the four alternatives to complete the corresponding

gap in the text. In the second parcel (vocabulary 2) the test taker is required to select the best possible choice to fill in the blanks at sentence level. With respect to the symmetrical formats of the two sections and the estimated weight of the knowledge of vocabulary in the UTEPT, it was indicated that the two sections significantly account for the performance on vocabulary section. Also, the 68% of factor loading of the whole section, reveals the high weight of vocabulary knowledge in language proficiency. Thus, the measures estimated in the analysis of the vocabulary section of the UTEPT are considered convincing as far as the internal consistency of the estimations is concerned.

The assessment of reading has been given a special attention in language testing (Alderson, 2000). In the UTEPT the reading module consisted of two separate sections. In the first section, multiple choice reading comprehension question items (reading 1) followed by texts and in the second section the restatement of the sentences as paraphrased structures (reading 2) are employed.

The results of the first-order model for reading comprehension revealed that the multiple choice reading comprehension question items parceled to account for their cumulative contribution to the first-order level of construct were working appropriately as the parceled indicator loaded at an acceptable level (β=.70). The number of items included in the parcel was found sufficient. The regression load of the parceled multiple choice question items was moderate and above, whereas, the loads estimated for the items in the restatement section was found inappropriate and was not qualified to account for an acceptable level of variation explained. However, the reading section in the UTEPT in a general model revealed an acceptable average load of 0.70%. The relatively smaller values estimated for the unexplained variances by the two associated error terms were negligible.

The estimations for the reading comprehension section in the UTEPT first-order level of measurement were used in order to examine their contribution to the fitness of the final model into the empirical data. The results, as appeared in Figure 2 and Table 4, revealed that the model was greatly influenced by the invited variance explanation from the reading comprehension section. The degree of improvement of the model is to a large extent influenced by the variance explained by the reading indicators included in the first-order model. However, the parceled items of restatement section have unfortunately influenced the model negatively ($\beta = 0.36$) and have contributed to the unexplained portion of the variance observed in the final model. It is estimated that the variation unexplained must be influenced by the fewer number of items parceled into the restatement section.

The present research aimed to contribute to the developments in language testing both practically and theoretically. From the theoretical perspectives, examining the application of SEM methodology in establishing a sound method of investigating the contribution of measurement models, testing these measurement models in a model of higher-order level of construct in a covariate matrix is assumed to be highly reliable. The method is self-sufficient as it does not need to develop separate models of correlation for every construct at their primary level to determine if the constructs are qualified to be entered into a higher-order model. Prior to the application of SEM methodology, and before the innovative ideas of constructing a moment analysis of covariance as it is performed in the SEM, it was impossible to judge the reliability of a system of measures all in one model. Furthermore, the developed model could be tested to ensure the theoretical model of the components of the test, e.g., the Bachman's Model of Language Proficiency (1990), could be directly tested from the empirical data. Yet, from another

perspective, SEM methodology can be used to investigate the widely supported data-driven validity check. This study would provide further scientific evidence to apply the method in future to other available large-stake language and cognitive tests.

From practical perspective, the question of validity of a test is claimed to be answered in terms of its repeated administrations. The richer the item bank, the easier it will be to provide validity evidence through quantitative methods. In order to do so, researchers need to frequently ensure that the fitness of the collected data across various administrations comply with the postulated theoretical framework underlying the test (Chapelle, Enright, & Jamieson, 2008, p. 61). In the absence of a tested framework, researchers may feel the need to re-run the factor structure studies every time the test is administered in order to refresh the item bank and ensure the route has been maintained across the repeated administrations.

The present research was the first attempt to analyze UTEPT with a SEM-based framework. The framework appears to be adequate in providing convincing insights for the latent trait of the test. It seems that it is useful for the examination of these traits in terms of the test-takers characteristics and the academic reforms leading to different kinds of modifications.

References

Alderson, J. C. (1981). Report of the discussion on general language proficiency, In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing,* ELT Documents 111. London: The British Council.

Alderson, J. C. (1986). Innovations in language testing. In M. Portal (Ed.), *Innovations in Language Testing.* Windsor: NFRE-Nelson.

Alderson, J. C. (1991). Language testing in the 1990s: How far have we come? How much further have we to go?. In J. C. Anivan (Ed.), *Current developments in language testing* (pp. 1-19). Singapore: SEAMEO Regional Language Center.

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

Anivan, S. (Ed.). (1991).*Current developments in language testing* (pp. 1-19). Singapore: SEAMEO Regional Language Center.

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

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.

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

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

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: Testing factorial across two groups of children in the Korean/English two-way immersion program. *Language Testing*, *15*(3), 380-414.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modelling with small samples: Test statistics, *Multivariate Behavioral Research, 34,* 181–197.

Byrne, B. M. (2010). *Structural equation modelling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Routledge: Taylor & Francis Group.

Canale, M., & Swain, M. (1980).Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Celce-Murcia, M., Dornyei, M. Z., & Thurrell, S. (1995). Communicative competence: A pedagogical motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2), 5-35.

Carroll, J. B. (1993). *Human cognitive ability: A survey of factor analytic studies.* Cambridge University Press.

Chapelle, C., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language.* New York: Routledge.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement*. Avon, England: Multilingual Matters Ltd.

Eckes, T., & Grotjahn, R. (2006).A closer look at the construct validity of C-tests. *Language Testing*, *23* (3), 290-325.

ETS. (2002). *LanguEdge courseware score interpretation guide.* Princeton, NJ:

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), *Issues in language testing research* (pp.11-29). Rowley, Mass: NewburyHouse

Farhady, H. (2005). Language assessment: A Linguametric perspective. *System*, *2*(2), 147-164.

Farhady, H., & Abbassian, G. R. (2000). Test method, level of language proficiency and the underlying structure of language ability. *Alzahra Journal, 9*(29), 27-32.

Hoe, S. L. (2008). Issues and procedures in adopting structural equation modeling technique. *Journal of Applied Quantitative Methods*, *3*(1), 76-83.

Hughes, A., & Porter, D. (1983). *Current developments in language testing.* UK: University of Reading, 3-37.

In'nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing*, *29*(1), 131–152.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach*. Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Milanovic, M. (1988). *The construct validation of a performance based battery of English language progress test.* Unpublished PhD dissertation, University of London.

Oller, J. W., Jr. (1978). The language factor in the evaluation of bilingual education. In J. Alatis (Ed.), *International dimension of bilingual education.* Washington, D. C.: Georgetown University Press.

Oller, J. W., Jr. (1983). A consensus for the eighties?. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 351–356). Rowley, MA: Newbury House.

Oller, J. W. Jr., & Damico, J. S. (1991). Theoretical considerations in the assessment of LEP students. In E. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students*. Austin: Pro-ed publications.

Oller, J. W., Jr., & Hinofotis, F. A. (1980). Two mutually exclusive hypotheses about second language ability: Factor analytic studies of a variety of language subtests. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing* (pp. 13–23). Rowley, MA: Newbury House.

Pasternak, M., & Bailey, K. M. (2004). Preparing nonnative and native english-speaking teachers: Issues of professionalism and proficiency. In L. D. Kamhi-Stein (Ed.), *Learning and teaching from experience: Perspectives on nonnative English-speaking professionals* (pp. 1 55-175). Ann Arbor: The University of Michigan Press.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York, NY: Peter LangPublishing, Inc.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-Based test. *Language Testing, 26*(1), 005- 030.

Schroeders, U., Wilhelm, O., & Bucholtz, N. (2010). Reading, listening, and viewing comprehension in English as a foreign language: One or more constructs? *Intelligence*, *38*, 562–573.

Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*(1), 31-57.

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modelling approach. *Language Testing*, *25*(1), 435-464.

Spolsky, B. (1973). What does it mean to know a language, or how do you get someone to perform his competence? In J. W. Oller & J.C. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher*, Rowley, Mass.: Newbury House.

Stricker, L. J., Rock, D. A., & Lee, Y. (2005). *Factor structure of the languedge™ test across language groups*, ETS, Princeton, NJ.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed). Pearson: Allyn and Bacon.

Vecchio, A. D., & Guerrero, M. (1995). *Handbook of English language proficiency tests*. EAC-West, New Mexico Highlands University, Albuquerque.

Vollmer, H. J. (1983). The structure of foreign language competence. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp.3-30). UK: University of Reading.

Vollmer, H. J., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In Oller, J. W., Jr. (Ed.), *Issues in language testing research* (pp. 29–79). Rowley, MA: Newbury House.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*(1), 119 - 140.