Available at jtpsol.journals.ikiu.ac.ir

Journal of Teaching Persian to Speakers of Other Languages

Corpus-Based Insights into Modeling a Level-Specific Persian Language Proficiency Test (PLPT): Development and Factor Structure of the PLPT Listening Tasks

Mahmood Bijankhan

Full Professor, University of Tehran

Parvaneh Shayestefar*

Post-doctoral Researcher, University of Tehran

ABSTRACT

The factor structure of the listening section of a Persian Language Proficiency Test (PLPT), developed and used for academic purposes, was examined in this study. A Structural Equation Modeling (SEM) was employed using AMOS (V. 18) to analyze the responses of a number of Persian language learners (n=120) who participated in the first piloting phase of the test in 2014. To examine whether the listening factor corresponds to the test hypothesized structure, three models (unitary, correlated and uncorrelated) were postulated on the basis of the literature. The results from model testing suggested that the correlated model (i.e., correlated receptive skills of listening and reading) fitted the obtained data best, supporting the reporting of distinctive listening and reading factors. The results of the current pilot study provide empirical evidence for reporting valid listening scores and interpretations based on separate scores found for the PLPT listening skill. Implications for Persian language teaching, learning, and assessment are discussed.

Keywords: Persian Language Proficiency Test (PLPT), Academic Version (AV), Persian language learners, Written Corpus

* Received on: 2/2/2016 Accepted on: 21/6/2016

E-mail: shayestefar.parva@ut.ac.ir
parishayeste@yahoo.com

2. Introduction

Corpus linguistic research demonstrates the benefits of using corpora in language teaching. Alongside the importance of corpus linguistics (CL) for the development of reference materials, resources, and approaches that can be used in classroom teaching, there is an emerging recognition in the use of corpora in language testing and assessment (LTA). Since corpora include collection of authentic texts of naturally occurring discourses (written or spoken), test designers and researchers have been investigating, discussing and evaluating the ways in which CL can provide evidence for the development of realistic tasks and rating criteria. These worthy evidence and information have placed language test designers in a privileged position, with theoretical and empirical description of authentic language for language assessment. In other words, test developers have increasingly used CL as a reference resource to identify the linguistic characteristics of the native speakers' usage suggesting aspects of language to test or to avoid (Park, 2014). Such data are used to describe what both learners and proficient users can do at various proficiency levels in a particular language, hence, making it possible to assess the linguistic features found in learners' language against those associated with native users of that language. Despite the growing recognition and availability of native corpora for LTA, few test makers have applied them in developing and validating language proficiency. To narrow this gap, initiatives were taken by the present study.

Test developers have now turned to corpora to develop and validate specific levels of language proficiency ranging from the lowest to the highest proficiency levels. This implies that corpus evidence informs the development of a list of linguistic indicators used to differentiate learners of different proficiency levels from each other or from the native speakers (Barker, 2010). Typical performance indicators associated with proficiency levels have been developed by the Association of Language Testers in Europe (ALTE) in terms of a set of 'Can-do' statements. These statements are aligned to the proficiency levels described by Common European

Framework of Reference (CEFR; Council of Europe, 2001). Grounded in the theories of communicative competence (e.g., Canale & Swain, 1981; Davies, 1989; Hymes, 1972), CEFR provides a coherent framework for description of learners' communicative language ability (CLA) at each of the six levels (i.e., A1 to C2), such as *can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated* (A2 level listening overall general ability; see CEFR, 2001; and ALTE) or *Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail* (B2 level Speaking overall general ability; see CEFR, 2001; and ALTE).

On the basis of such usage-based characteristics of proficiency levels, CEFR, thus, identify linguistic features that are 'criterial' for distinguishing one proficiency (i.e., communicative competence) level from the others. The criterial features which are typically found across all CEFR six levels are *lexical, grammatical, semantic, phonological, orthographic, sociolinguistic* and *pragmatic*. In CL, these features are also described as linguistic features that are especially common in academic corpora, and taken as important indicators of differences among discourses and registers. Although many test developers are unsure about how to use the information in the CEFR framework to design tests that are aligned with the framework (Harsch & Rupp, 2011), when modeling and validating the proficiency tests are considered, the framework has appeared much influential in LTA (see e.g., Milanovic, 2009; Taylor & Jones, 2006). The challenge, however, has been to develop tests that would be a reliable and valid measurement of language ability. Concerns for such a challenge have resulted into the development of a framework such as CEFR which aims to provide descriptions of proficiency levels. Such a critical need is also highlighted by Bachman and Palmer (2010), acknowledging that for the development of any standard test such as proficiency tests, language test developers need to select a framework that describes attributes of language users that are involved in language use. In

this sense, the framework provides an appropriate tool to measure communicative knowledge and ways of using this knowledge (Spolsky, 1989).

The present work is situated within such a broad context and aims to provide insight into the question of how development and validation of a Persian Language Proficiency Test- Academic Version (PLPT-AV) is informed by a corpus of written Persian called *Peykareh (*a 100-million-word corpus developed by Bijankhan, Sheykhzadegan, Bahrani and Ghayoomi, 2011). We focus on Persian as a foreign language of non-Persian applicants of Iranian universities. To support reliable assessment of language performance of these learners, Peykareh was analyzed for its criterial features. Tests for all four skills (listening, reading, writing and speaking) were developed during the PLPT project that was sponsored by the Ministry of Science, Research and Technology (MSRT) of Iran; however, the present study focuses on the development process of the PLPT-AV listening section that was made aligned to the major proficiency levels (CEFR levels: basic users; independent users; and proficient users). Notwithstanding valuable endeavors made to develop Persian proficiency tests, for example, a Persian proficiency test designed by Ghonsooli (2010) and theoretical foundations of PLPT-AV provided by Sahraie and Jalili (2012), no evidence supported the existence of a standardized leveled test of Persian prior to 2015.

Given the recent initiatives in the development of the PLPT-AV, little is known about its validation, especially the factor structure of its receptive skills (listening and reading sections). This study investigates the factor structure of the PLPT-AV listening section using Structural Equation Modeling (SEM). Based on the available validation studies on proficiency tests, it will be of an interest to examine whether the PLPT-AV listening module is a valid test for measuring test-takers' listening ability.

3. Literature Review

An increasing direct involvement of corpora in LTA occurred with systematic electronic collections of written and spoken data by institutions

and examination boards. In this regard, large representative corpora, native corpora, learner corpora, and specialized corpora have been actively used in developing and validating language tests. For instance, the International Corpus of Learner English (ICLE) developed at Center for English Corpus Linguistics (CECL, in Belgium) was set up around 2000's for such purposes. Having included argumentative essays and literature papers collected by research teams worldwide, ICLE was aimed to support the Contrastive Interlanguage Analysis approach of its developers (Granger, Hung & Petch-Tyson, 2002). During the 1990s, the EFL Division of the University of Cambridge Local Examinations Syndicate (UCLES EFL) and Cambridge University Press developed Cambridge Learner Corpus (CLC) as a unique archive of learning writing scripts, demographic and score data. CLC initially included three proficiency levels of general English tests (i.e., the First Certificate in English (FCE), Certificate in Advanced English, and Certificate of Proficiency in English), however, it expanded to include other domains and proficiency levels beyond these three English exams.

The implications of corpora in LTA have been acknowledged by research over the past decades. Alderson (1996) suggested the use of corpora in test compilation and selection, test preparation, scoring, calculation and delivery of test results. Likewise, Barker (2004) maintains that both native and learner corpora can reveal much about developing test materials, publishing and assessing. As an example, a corpus-based checklist was developed in the U.K. to validate academic IELTS speaking tests in terms of communicative functions in different domains (e.g., Brooks, 2001). Moreover, Hughes (2008) explored the impact of edited authentic texts on the language within thirteen reading passages of FCE. Comparing edited with the original versions in particular comparing their lexis with native corpus frequencies, he asserts that these reading tasks provide phraseologies that test takers normally expect to meet in real-world language.

In test designing process, corpora can help language test designers to identify and apply Reference Level Descriptions (RLDs) across all proficiency levels. As an example, the thirty-million-word CLC was used to

provide a set of RLDs for English for all six levels of CEFR from A1 to C2. Examples of the criterial features established for distinguishing one CEFR level from the others are *lexical semantic, morpho-syntactic, syntactic, and pragmatic features*. These features will result into a valuable data source about the nature of language proficiency at all the CEFR levels. Corpora are explored and analyzed to show which collocational patterning of these features are frequent or less common at particular proficiency levels suggesting what learners can be expected to know at these levels. In Hawkey and Barker's (2004) perspective, frequency of occurrence obtained through corpus analysis shows how to use CL to inform test life-cycle and validation procedures. Such a corpus-based frequency information can add both "grammatical and lexical details to CEFR's functional characterizations of different levels" (Hawkins & Filipovic, 2012, p.5), also quantify the criterial features or RLDs needed to distinguish between the six levels of CLA of the CEFR framework.

Though the available literature provides evidence suggesting that CL and corpus-driven approaches have practical implications for LTA, the literature is not yet developed on high-stakes PLPT. The present study, therefore, represents an initial attempt to develop and construct-validate a newly designed PLPT-Academic version. This test has been aligned with the communicative competence framework to assess communicative language abilities of non-Persian university students. To the best of our knowledge, there are no reports on a Persian proficiency test using CL data for linking the test-takers' abilities to certain levels of proficiency. To understand how the scores of the corpus-based PLPT-AV relate to the listening construct being measured, the factor structure of its listening section was investigated in this study. The results then would provide the appropriateness of using separate listening scores to report the validity of componential PLPT-AV. In what follows, we describe the methodologies adopted for developing the PLPT-AV and assessing the factor structure of its listening section.

2.1. Development of the PLPT-AV

The present project was sponsored by MSRT with the ultimate aim of developing a standard language proficiency test in Persian. The study phase of this project was initiated in 2013, and a year later, the development phase was set up with triangulated sources of data from Persian language policy documents, Peykareh Corpus, available materials on Teaching Persian to the Speakers of Other Languages (TPSOL) or what is locally called as Amooszesh-e Zaban-e FARSI (AZFAN), AZFAN or TPSOL instructors, and language test experts. The outcome of these two phases led into the skills (reading, listening, writing and speaking) and tasks specifications to elicit the desired information through the PLPT-AV framework.

For the purpose of a leveled-test design, it is essential to define the domain of knowledge and skills, relevant content, and the way that content is assessed through the test. On such basis, the present project involved decisions on content coverage and content representativeness of the PLPT-AV tasks or items, performance standards in terms of scaling descriptors, i.e., 'Can-do scales of successive levels of proficiency', types and numbers of the tasks, relative weights of skills and tasks, item specifications (detailing about item acceptable vocabulary, syntax and content limits, item numbers, …) together with the scoring/rating procedures. The point of departure for this process was CLA model (Bachman, 1990) taken as an integrated model of *Linguistic Competence*, *Sociolinguistic Competence* and *Pragmatic Competence*. Each competence is described with a set of relevant 'criterial features' or 'Can-do statements'. For instance, '*Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment)'* (basic users: A1), or '*Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of Proficient meaning even in more complex situations'* (proficient users: C2). The scale underlying these features or descriptors was divided into six levels (A1, A2, B1, B2, C1, C2) that allow for functional communication, i.e., CLA manifestations, ranging from limited (A2) to successful (C1, C2) use of Persian language.

The PLPT-AV included both receptive and productive/interactive activities and strategies. The former included Multiple Choice (MC) items (6 tasks in listening; 6 tasks in reading) and the latter involved written interaction (3 tasks), and face-to-face and interaction strategies (2 tasks in Speaking). Three professional TPSOL experts: two from Persian Language and Literature department of University of Tehran, and one from AZFAN Institute of University of Tehran (i.e., Dehkhoda Lexicon Institute and International Center for Persian Studies), one computational linguist and three postgraduate students of linguistics assisted the present researchers. The corpus-extracted resources we employed as 'criterial features' to inform the specifications of tasks and items come below.

2.2. Peykareh: A Written Corpus of Contemporary Persian

Built as a written language resource for the contemporary Persian, *Peykareh* (see Bijankhan et al., 2011) is a large corpus (100 million words, 35058 texts) designed at the Research Center for Intelligent Signal Processing (RCISP) The corpus texts collected from naturally occurring discourse of different academic, institutional or constitutional registers (e.g., Education, Manuals, Regulations, Conversation…) include 'texts written to be read' (WR, 87%) and 'texts written to be spoken' (WS, 13%). Peykareh has been searched closely through Searchdata tool to look for its syntactic and morphological resources. The outcome, according to Bijankhan and his associates (2011), was the emergence of more than a dozen of general parameters: *relative*, *complement*, *conditional*, *subordinate*, and *passive structures, question words constructions, noun*, *verb*, *adjective*, *adverb*, *preposition* and *pronoun constructions, articles, homographs/homophones*, all in forms of monograms or strings of words of collocated bigrams, trigrams, or in general, *n*-grams. The frequency of occurrence of each parameter was used as a criterion for selection and inclusion of each parameter within the test tasks and items contents, with highest frequencies for A1, A2 levels, the moderate for B1, B2, and the least frequencies for C1, C2 levels.

To determine a cut score, a group of experts (judges) are required to take part in the standard setting process to define a cut score for a certain test (Kollias, 2012). On such a base, the policy committee of the present project (the project researchers, two experts from MSRT and two AZFAN experts) reviewed the recommended scores and made the final decisions on the cut score. The results came into a total scale of 0 to 120 points, with each section (Reading, Listening, Speaking and Writing) receiving a scaled score from 0-30 (A2-1=0-5; A2-2=6-10; B1=11-15; B2=16-20; C1=21-25; C2=26-30). For instance, a cut-off score of 10 or lower says that numerical score of 10 or lower grants a particular level or lower than the level of the cut-off score (e.g., A2), while the numerical score of 10^{+} points to a particular level (e.g., B1) or higher. Both listening and reading sections of the PLPT-AC were rated on a 0-30 scaled score. Each task of these two receptive skills included 5 MC items.

What followed tasks and item specifications as well as score definition was checking on the clarity and comprehensibility of the PLPT-AV questions and rubrics also to estimate the required time. For such a purpose, two AZFAN experts and one Persian language learner who had been learning Persian in Iran for more than 15 months reviewed the tasks and items. The outcome was the elimination of problematic items such as perceived ambiguous items (3 items) and complex structures with multi-unit constructions (e.g., 5-6 collocated strings) that could not fit the level-specific indicators.

In pretesting the test to a sample of 30 foreign students with different L1 (in February 2015), information on psychometric characteristics were indicated. Cronbach's alpha consistency was estimated and the reliability coefficient was found to be .82. More importantly, the PLPT-AV needs to be examined in terms of the factor structure of its receptive construct. Possible models of factor structure of the PLPT-AV are focused upon below.

2.3. The Hypothesized Factor Structure Models

The factor structure of the PLPT-AV can be hypothesized through the modality of its output and the reporting format of its scores to the examinees.

Along with a single total score, separate scores for listening and other sections are reported to the PLPL-AV examinees. Following In'nami and Koizumi's (2011) conclusion that a single total score shows a higher-order factor underlying test performance, separate listening and reading scores were used to hypothesize that distinctive factors underlie performance on both sections. Therefore, a correlated trait model was hypothesized to show two separate skills involved in the total receptive skill. Such a structure concurs with the literature on language ability as a trait consisting of underlying correlated specific abilities (e.g., Bachman & Palmer, 1981; Sang, Schmitz, Vollmer, Baumert & Roeder, 1986). Nevertheless, a single trait model of language ability has been also reported by Oller's (1983) empirical study on students' performance on L2 Placement Examinations of the University of California. Having analyzed 164 students' responses on these tests, he reported L2 language ability as a single trait. In another stance, Wilson (2000) believes that listening ability is not correlated with other language abilities such as speaking or reading. Aligned with these inconsistent views, it was hypothesized that listening ability of the PLPLT-AV is a) inseparable from reading (i.e., unitary model), b) separable but uncorrelated with reading; or c) separable but correlated with reading.
See Figures 1, 2 and 3, below.

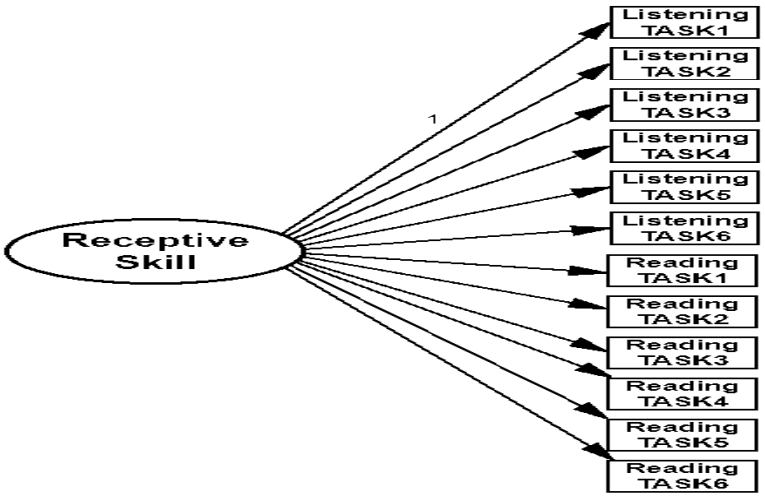

Figure 1. Unitary model

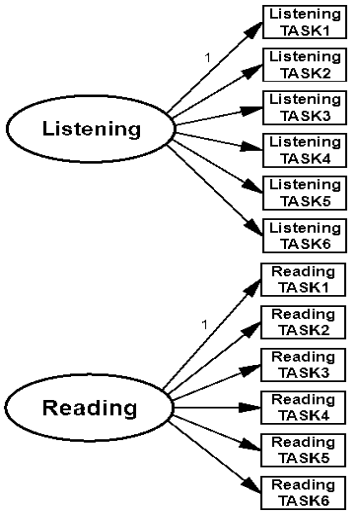

Figure 2. Uncorrelated model

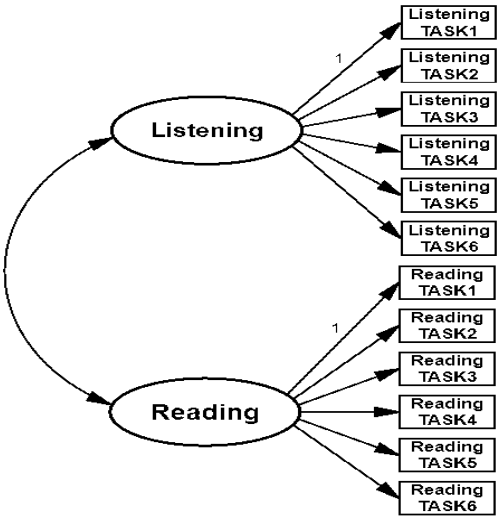

Figure 3. Correlated model

2.4. Research Questions

This study aimed at examining whether test-takers' observed data support the converging evidence of the componential model of Persian receptive skills and whether the separate nature of receptive skill of listening is supported. The following research question was investigated herein:

Does the listening component of the PLPT-AV correspond to the hypothesized factor structure underlying the test?

In other words,

Taking into account the PLPT-AV listening skill, does the assumed correlated model fit the data better than the uncorrelated and unitary models?

4. Method

3.1. Participants

A total number of 120 participants from 16 European and Asian countries participated in this study. Their average age was 28.5 years, ranging from 19 to 51. The sample consisted of 81 (67%) males and 39 (33%) females who had enrolled in either an undergraduate or graduate programmes in University of Tehran or its institute of Persian language learning (called Dehkhoda Institute) during the academic years of 2013-2015. The largest

groups were from Iran's neighboring countries, i.e., Asian countries (75%), and only 24% of participants were from European countries.

The test performance data were obtained from the newly designed PLPT-AC tests administered among the participants, in consultation with authorities of University of Tehran. Care was taken in the administration: test booklets and materials were securely guarded, audio systems were checked along with the physical setting, and proctors received short training. This included information about administration timetable, administration guidelines (whether to admit latecomers, how to behave during the test, guiding test-takers through their seats, etc.), delivering the test booklets to the test-takers, and returning them back to the researchers.

3.2. Instruments and Procedures

The database used for this study was part of a larger PLPT development project sponsored by MSRT during 2013-2015. The PLPT-AV assesses different skill domains, including four independent domains (listening, reading, speaking and writing). Receptive skill of listening was defined as an integrated-skill domain including listening/reading tasks, and expressive skill of speaking, though not the focus of the present study, was defined as integrated reading/speaking domain. Although such shared-skill domains can be used as individual latent variables (Pae, 2012), decision was made based on the modality of the output. Therefore, listening skill was instrumental for a listening comprehension output.

The PLPP-AV was used for measuring the overall language proficiency of Persian language learners. The construction, standard setting, and modes of the tests were determined by the test development committee at University of Tehran. Five AZFAN experts including linguists and researchers were consulted for their expertise, ideas and critical views. The outcome was a 120-point PLPT-AV measuring all four Persian language skills through four 30-point measures.

Given the CLA framework for the development of the listening skill, 6 communicative events were specified (e.g., asking for services, radio interviews, lectures, etc.,), checked for length and difficulty levels, and

finally recorded in forms of dialogue, conversation, and monologue tasks. Each task was followed by 5 independent MC items that were objectively marked. Regarding the length and difficulty, the tasks were not equivalent; they were designed and sequenced in an increasing length and difficulty level requiring both linguistic and non-linguistic knowledge to work interactively to produce comprehension. Overall, the listening booklet included 30 items for measuring different skills such as scanning, perceiving the utterance, identifying and understanding the message, and interpreting the message.

The 6 listening tasks, paced by a compact disk and tape-recorder, lasted for 50 minutes. Ten extra minutes were given at the end to let test-takers transfer their answers from their listening booklets to a separate answer sheet. The present researchers were present at the time of test administration, assisted in test administration, and finally scored listening and reading papers based on a key-answer sheet designed during the test development process. The total mean score of the listening performance was found 21.43 (X=18.70 for reading skill). The Cronbach's Alpha reliability coefficient was 0.87 for the PLPT-AV listening test.

3.3. Analysis

The data were in form of the scores for listening and reading tasks measuring the ability to a) locate straightforward factual information; b) infer gist and purpose of short spoken texts based on explicit information; c) infer gist and purpose of extended spoken texts based on explicit information; d) understand details in short spoken texts, e) understand details in extended spoken texts; and f) understand and interpret critically all forms of written language including abstract, complex texts of implicit or explicit meaning. Both the scores and the percentage of correct responses in each subskill were available for the analysis. The subskill items were used as measures of listening and reading constructs in the present study (see Figures 1, 2 & 3), thus, their scores were used for observed variables in each model. SEM analysis was employed to examine the factor structure of the PLPT-AV listening skill. Maximum likelihood method was used for the purpose of

estimating model parameters. Kurtosis and Skewness values were checked for the normality of distribution of the variables.

4. Results

4.1. Descriptive Statistics

The descriptive statistics in Table 1 show that skewness and kurtosis values of the items are within |3.30| (*z* score at $p<.01$), suggesting no violation to the univariate normality assumptions. Mardia's coefficient was also checked for multivariate normality and the obtained value was below the recommended value of 20.00 (Harington, 2009) thus indicating multivariate normality of the data.

Table 1.

Descriptive Statistics for Subskills of PLPT-AV

	N	Minimum	Maximum	Mean	Std. Deviation	Kurtosis	Skewness
Listening 1	120	9.00	17.00	13.14	1.55	-.322	.291
Listening 2	120	10.00	18.00	13.67	1.53	.375	.048
Listening 3	120	9.00	16.00	13.65	1.24	-.767	2.415
Listening 4	120	6.00	21.00	12.16	2.12	.314	1.981
Listening 5	120	7.00	20.00	12.57	1.66	.671	3.298
Listening 6	120	3.00	19.00	13.17	2.34	-1.028	3.293
Reading 1	120	8.00	18.00	15.12	1.64	-1.706	3.481
Reading 2	120	8.00	17.00	12.16	1.77	-.132	.253
Reading 3	120	9.00	18.10	13.93	1.83	-.330	.027
Reading 4	120	4.00	16.00	11.41	1.85	-.616	1.984
Reading 5	120	6.00	19.00	12.31	2.17	.086	1.416
Reading 6	120	1.00	26.00	12.71	4.03	1.017	1.578

4.2. SEM Analysis: Testing the Three Hypothesized Models

The study tested the extent to which the three hypothesized models of the test components are consistent with the obtained data. Table 2 shows the model fit indices obtained from running AMOS analysis. The results in Table 2 indicate that the chi-square statistic ($\chi 2=89.97$), degrees of freedom (DF=49), normed Chi-square (CMIN/DF =1.84), Root Mean Square Error of Approximation (RMSEA=.084), Goodness-of-Fit Index (GFI=.88) and Akaik Information Criterion (AIC=147.97) of the unitary models are more

acceptable than those of the uncorrelated Model ($\chi2$=102.59; df=55; RMSEA=.085; GFI=.87; AIC=148.59).

Table 2.

Goodness-Of-Fit Indices of the Three Hypothesized Models

Model	**Fit statistics**						
	$\chi2$	*df*	CMIN/DF (1>, 3<)	GFI (≥.90)	RMSEA (≤.08)	AIC (the lower)	BIC (the lower)
Unitary	89.97	49	1.84	.88	.084	147.97	228.81
Uncorrelated	102.59	55	1.86	.87	.085	148.59	218.18
Correlated	64.84	38	1.71	.92	.070	144.83	218.18

Note. df=degrees of freedom, $\chi2$=Chi-Square

For the latter, only one comparison statistics, i.e., Bayesian Information Criterion (BIC) was found more acceptable than that of the unitary model. Despite this, when GFI and RMSEA values are considered none of these models seems appropriate. Nevertheless, Table 2 shows that the Goodness-of-Fit indices of the correlated model were better than those of the uncorrelated and unitary models, so the correlated model was the best model for the present data, showing an interpretable and meaningful model for Persian language receptive skill of listening. The model factor loadings were statistically significant ($p<.05$), ranging from .40 to .79 for the listening subskills (see Table 3). However, only three factor loadings appeared significant for the unitary model. Overall, the results show that the unitary and uncorrelated models were less favorable than the correlated model. The path coefficients in Figure 4 were significant at the alpha level of .05 and all Standard Errors (S.E.) were smaller than 1.0 indicating no violation of estimates. The analysis of the items loadings on the two latent factors, factor loadings, and p- values of the correlated model could well support the construct of listening, i.e., the correlated model with two separate factors of listening and reading was confirmed. The path coefficients of the observed variables (ListeningTASK1 to ListeningTASK6) to the corresponding factor of listening were moderate to high (from .40 to .79) and the correlation between the listening and reading factors was acceptable (almost .70), though less than .90, therefore the factor of listening can be considered significantly distinct from reading construct.

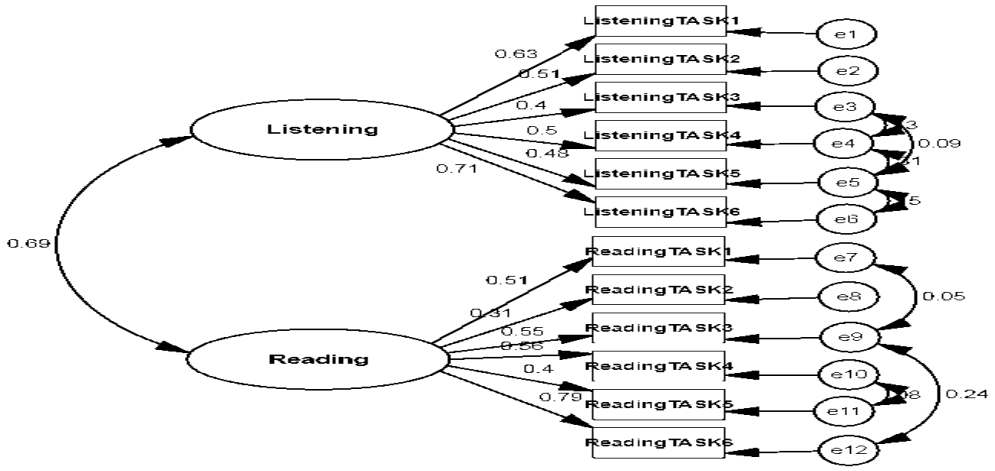

Figure 4. Final model with standardized factor loading indices

Table 3.

Results for the Measurement Model (Correlated Model)

			Estimate	.E.	.R.	P
ListeningTASK1	---	Listening	.630			
ListeningTASK2	---	Listening	.508	199	3.914	**
ListeningTASK3	---	Listening	.401	154	.486	013
ListeningTASK4	---	Listening	.489	408	2.525	012
ListeningTASK5	---	Listening	.477	412	1.927	050
ListeningTASK6	---	Listening	.709	372	.491	**
ReadingTASK6	---	Reading	.788	554	6.059	**
ReadingTASK5	---	Reading	.398	256	3.419	**
ReadingTASK1	---	Reading	.506			
ReadingTASK2	---	Reading	.297	221	2.268	023
ReadingTASK3	---	Reading	.551	265	.924	**
ReadingTASK4	---	Reading	.552	225	.673	**

Note. Estimate=Standardized factor loadings; P=p-value<.05; S.E.<1.00; Critical Ratio (C.R.)>1.96

5. Discussion and Conclusion

The assessment of listening and reading skills has received considerable attention in LTA (Alderson & Banerji, 2001). Aligned with this, the question of validity of these two test constructs has formed a significant focus of language testing discipline (Alavi, Kaivanpanah, & Nayernia, 2011; Alderson et al., 1995; Bachman, 2004; Bachman & Palmer, 1996). The present study aimed to examine the underlying factor structure of the listening construct of the PLPT-AV, a corpus-based Persian proficiency test that was aligned to proficiency levels. In order to investigate empirically if the underlying listening factor of the PLPT-AV corresponds to the proposed theoretical models of language proficiency arguing for underlying separable but correlated specific abilities (Bachman, 1990; Bachman & Palmer, 1981; Sang, Schmitz, Vollmer, Baumert & Roeder, 1986), a confirmatory factor analysis was used on a sample of Persian language learners from different Asian and European countries. The three competing models for receptive skills hypothesized based on the literature were examined to see which fit the data better. The results of the measurement and structural models of correlated constructs of listening and reading support the construct of listening as measured in the ALPT-AV. The correlated model (Figures 3 & 4) was confirmed against the obtained data, with the 6 listening tasks loading satisfactorily into it. The regression loads of these communicatively-based questions appeared moderate to moderately high and high, whereas the loads estimated for the listening factor in both uncorrelated and unitary models were not satisfactory enough to account for an acceptable level of variation explained by the models. The results broadly support the current reporting of two scores corresponding to the receptive modalities.

Such an empirical support for the correlated model of listening and reading factors of the test corroborates results of factor structure of L2 language ability reported by Bachman and Palmer (1981). The fact that listening was found to be correlated moderately high with reading, though not highly correlated, suggests that listening is separable from reading but is a similar skill when the overall receptive skill is concerned. This finding is

congruent with the available evidence showing language ability as multi-componential (e.g., Sasaki, 1996; Shin, 2005). The loadings of each task on the listening modality showed that the listening module of the PLPT-AV can measure the listening sub-skills in testees' ability to listen to and comprehend the six listening tasks designed to measure their listening ability.

The evidence in support of the correlated factor structure of the receptive modalities of the PLPT-AV is also consistent with uni-dimensionality for listening comprehension reported by Wilson (2000) but inconsistent with bi-dimensionality for reading comprehension found by him. The unitary listening and reading factors observed to be correlated in the present study were found uncorrelated by Wilson (2000) Apart from the language, content and format of the PLPT, one possible explanation lies in the designing sources of the tests. The present language proficiency test has been structured on Peykareh's real texts produced by Persian speakers in real contexts. The parameters extracted from authentic texts of such a large corpus were purposefully used as RLDs across different levels of Persian language ability. Used for measuring listening ability, such linguistic parameters might have contributed to producing the uni-dimensionality of this skill. This, in turn, reflects that the listening section items are not psychometrically distinct from each other, a finding that is similar to the results of studies on the TOEFL test (e.g., Hale, Stansfield, Rock, Hicks, Butler & Oller, 1988; Schedl, Gordon & Tang, 1996) as an international standardized test of proficiency.

Therefore the PLPT-AV listening module consisted of separate sections/measures structured in an increasing difficulty level form. All sub-sections reflect communicative functions of Persian language, ranging from frequently occurring small talks to semi-formal conversations to the more formal speeches and lectures. The results of the correlated structural model revealed that the MC listening comprehension tasks, accounting for their cumulative contribution to the separate factor of listening, can work appropriately as the distinct indicators of the listening factor.

Overall, the present results concur with some current reporting of the application of the CP to test design and development where the integrated tasks contribute to the scores for the target constructs (e.g., Biber et al., 2004; Kennedy & Thorp, 2007). Evidence of what language users can do gives a way to the use of such CL-based data to describe typical abilities common to each proficiency level. More specifically, CL evidence can help test designers to take insights into deciding on specific constructs (i.e., listening) to be tested, writing realistic tasks corresponding to specific proficiency levels, setting realistic criteria to measure what a learner can already do or need to learn in order to achieve mastery of a particular ability level, and validating test constructs and their representative tasks against the real-life texts of various functions.

6. Implications and Limitations

The present findings regarding the factor structure of a corpus-based Persian language proficiency test has implications for LTA, both theoretically and practically. On a theoretical level, the presence of distinctive listening skill in the PLPT-AV supports the reporting of separate language ability skills. In other words, a relatively acceptable correlation between these two factors suggests the arguments of the distinct but related nature of language ability skills. Besides, the application of moment analysis of covariance as it is performed in the SEM methodology makes it possible to judge the reliability and plausibility of the theoretical model of language proficiency and its components.

The results also can help capture the CL state-of-the-art in terms of how CL can inform the development and designing of language proficiency tests. As such, this study would provide evidence to promote application of corpora in future to other large-scale language tests. From practical perspective, the application of corpora to test material design and development has a washback effect on how language communicative tasks are designed and how they influence Persian language teaching, learning and testing. This study demonstrates the usefulness of using corpora for realistic

task specification and content both for testing and teaching in AZFAN classes. The authenticity of the format and content of the CL-based tests materials can significantly influence AZFAN practitioners' adoption of real-life texts of Persian corpora that are used by PLPT test designers. Alongside this, aspects of communicative language abilities underlying the PLPT tasks are underscored by AZFAN teachers, consequently, paid attention to by Persian language learners.

However, there are some limitations that should be noted. First, although the present sample consisted of diverse Persian language learners, it was too small for investigating the factor structure of a new proficiency test. Thus, the results are not fully generalizable to the intended PLPT test-taking population. The confirmed model as found in this study needs to be replicated with larger samples of test-takers to see whether the confirmed factor structure will be supported in the other forms of the PLPT.

Acknowledgement

The authors would like to appreciate the 'Vice-chancellor for Research' and the 'Research Affairs Office' of the University of Tehran, for supporting the PLPT-AV project also for approving the project in form of a Post-Doctoral fellowship (No: 140/337231; March, 10, 2014) conducted in the Faculty of Literature and Humanities. The authors are also thankful to Dr. Darzi, the head of Dehkhoda Lexicon Institute and International Center for Persian Studies, and Mrs Azam-Sadat Navabi, the director of the Department of Education, from the same institute, for their timely arrangement and kind assistant with data collection from the institute. The institute's Persian Language Teachers' collaboration and care are appreciated, too.

References

Alavi, S.M., Kaivanpanah, SH. & Nayernia, A. (2011). The factor structure of a written English proficiency test. *Iranian Journal of Applied Linguistics*, *3*(2), 27-50.

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. A. Thomas and M. H. Short (Eds.), *Using corpora for language research* (pp. 284-259). London: Longman.

Alderson, J. C. & Banerjee, J. (2001). Language testing and assessment. *Language Teaching, 34*, 213-236.

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

ALTE. (2002). *The ALTE can do project. Articles and can do statements produced by the members of ALTE 1992-2002.* Retrieved from http://alte.org/downloads/index.php?doctypeid=10

Bachman, L. F. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press.

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

Bachman, L. F. & Palmer, A. S. (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.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Barker, F. (2004). *Corpora and language assessment: trends and prospects, research notes*. Cambridge: UCLES.

Barker, F. (2010). How can corpora be used in language testing? In Anne O'Keeffe & Michael McCarthy (Eds.), *the Routledge handbook of corpus linguistics* (pp. 633-646). New York: Taylor and Francis Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V. Cortes, V., Csomay, E. & Urzua, A. (2004). *Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus* (Publication No. RM-04–03), Supplemental Report No. TOEFLMS-25). Princeton, NJ: Educational Testing Service.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M. & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, *45*, 143-164.

Brooks, L. (2001). Converting an observation checklist for use with the IELTS speaking test. *Research Notes, 11*, 1-20.

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

Council of Europe. (2001). *Common European framework of reference for languages: Learning, Teaching, Assessment.* Strasbourg: Language Policy Unit.

Davies, A. (1989). Communicative competence as language use. *Applied Linguistics*, *10*(2), 157–170.

Ghonsooli, B. (2010). Development and validation of a PLPT. *Foreign Language Research, 57*, 115-129.

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer-learner corpora, second language acquisition, and foreign language teaching*. Philadelphia, PA: John Benjamins.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W., Jr. (1988). *Multiple-choice Cloze items and the Test of English as a Foreign Language*, TOEFL Research (Rep. 26), Princeton, NJ: ETS.

Harrington, D. (2009). *Confirmatory Factor Analysis*. New York: Oxford University Press.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1-33.

Hawkey, R. & Barker, F (2004) Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.

Hawkins, J.A. & L. Filipovic (2012). *Criterial Features in L2 English*. Cambridge: CUP.

Hughes, G. (2008). Text Organization Features in an FCE Reading Gapped Sentence Task. *Research Note, 31*, 26-31.

Hymes, D. (1972). On communicative competence. In J. Prides & J. Holmes (Eds.), *Sociolinguistics*: *Selected Readings* (pp. 269-293). Harmondsworth: Penguin.

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

Kennedy, C. & Thorp, D. (2007). A Corpus-based Investigation of Linguistic Responses to an IELTS Academic Writing Task, in L. Taylor and P. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment* (Studies in Language Testing vol. 19, pp. 316-77). Cambridge: UCLES and Cambridge University Press.

Kollias, C. (2012). *Standard Setting of the Basic Communication Certificate in English (BCCETM) Examination: Setting a Common European Framework of Reference (CEFR)*. Hellenic American University, Office for Language Assessment and Test Development (OLATD).

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes, 37*, 2-5.

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

Pae, H. K. (2012). *A model for receptive and expressive modalities in adult English learners' academic L2 skills*. Retrieved December 11, 2015, from Pearson Language Test. http://pearsonpte.com/research/ Documents/ ResearchNoteexpressivefinal2012-10-02GJ.pdf.

Park, B. (2014). Cognitive and affective processes in multimedia learning. *Learning and Instruction, 29*, 125-127.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J. & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing, 3*(1), 54-79.

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

Sahraei, R. M. & Jalili, S. A. (2012). Theoretical Foundations for development of a Persian Proficiency Test. *Research Notes of AZFA, 1*(1), 123-150.

Schedl, M. A., Gordon, P. C., & Tang, K. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items*. TOEFL Research (Rep. 53), Princeton, NJ: ETS.

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

Spolsky, B. (1989). *Conditions for second language learning: Introduction to a general theory*. Oxford: Oxford University Press.

Taylor, L. & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). *Research Notes, 24*, 2–5.

Wilson, K. M. (2000). *An exploratory dimensionality assessment of the TOEIC test (TOEIC Research Report).* (Publication No. RR-00-14), Princeton, NJ: Educational Testing Service.