Journal of Teaching Language Skills (JTLS)
36(4), Winter 2018, pp. 141-170- ISSN: 2008-8191
DOI: 10.22099/jtls.2018.27029.2372

On The Factor Structure (Invariance) of the PhD UEE Using Multigroup Confirmatory Factor Analysis

Hamdollah Ravand *
Assistant Professor
Vali-e-Asr University of Rafsanjan
ravand@vru.ac.ir

Gholam Reza Rohani
Assistant Professor
Vali-e-Asr University of Rafsanjan
rrohani@gmail.com

Fatemeh Faryabi
M.A.,
Vali-e-Asr University of Rafsanjan
f.faryabi1991@gmail.com

Abstract

The aim of the current study was twofold: (1) to explore the internal structure of the general English (GE) section of the university entrance examination for Ph.D applicants (PhD UEE)into the English programs at state universities in Iran, and (2) to examine the factor structure invariance of the test across two proficiency levels. Multigroup confirmatory factor analysis was used to analyze the responses of a random sample of participants (N=1009) who took the test in 2014. First, four models (unitary, uncorrelated, correlated and higher ordered) were estimated and compared to find the model that best represented the data. Then, the factor structure invariance of the test across the two proficiency levels was explored using multigroup confirmatory factor analysis. The higher-order and correlated three factor models showed the best fit to the data. The results also showed that the structure of the test remained invariant across both proficiency levels. The results supported the multi-componential view of language proficiency. It was found that there is no relationship between levels of language proficiency and the structure of the test. However, the results called into question the score-reporting policy for the PhD UEE and led to the conclusion that a single total score does not reflect the structure of the test.

Keywords: factor structure invariance, language proficiency, multigroup confirmatory factor analysis, university entrance examination

Received: 14/03/2017 Accepted: 06/01/2018
* Corresponding author

The university entrance examination for PhD applicants into the English programs at state universities in Iran (PhD. UEE) is a high-stakes test that is part of the procedure that screens applicants into English programs at PhD levels at Iranian universities. The test is developed and administered by the Measurement Organization (MO). The MO tests screen applicants into Iranian universities at all levels of education: Bachelor's, Master's, and PhD programs. There is a very tight competition among the applicants to find a seat in a PhD program at a state university in Iran. Finding a seat at a state university is almost tantamount to securing a future job for most of the Iranian applicants. Consequently, the construct validity of these national matriculation tests in general and the PhD UEE should be scrutinized. According to Bachman (2005), an essential aspect of building a validity argument for a test is to investigate the internal structure of the test to make sure the interpretations made based on the test results are warranted.

The PhD UEE is a multiple-choice test which is designed to measure candidates' content knowledge, scholastic aptitude, as well as their general English knowledge (GE). The GE section is intended to measure language proficiency of the test takers in the areas of grammar, vocabulary, and reading comprehension. For the present study, the factor structure of the GE part of the PhD UEE test was investigated.

Although some sporadic studies have been done on the validity of the UEE for Bachelor's and Master's applicants into the English programs in Iran (e.g., Barati, & Ahmadi, 2010; Ravand & Firoozi, 2016) few studies have been conducted on the construct validity of the PhD. UEE (e.g. Ahmadi et al., 2015; Alibakhshi & Ghandali, 2011) in general and invariance of its factor structure in particular. Some studies investigated the predictive validity of the UEE (Alavi, 2012; Jamalifar, Chalak, & Heidari, 2014). Some other studies investigated the washback effect of the UEE on the attitudes of the teachers toward the test and the teaching practices of the English teachers at schools (Mahmoudi & Bakar, 2013; Salehi, & Yunus, 2012). Other studies (e.g., Barati, & Ahmadi, 2010;

Barati, Ketabi, & Ahmadi, 2006; Birjandi, & Amini, 2007; Firoozi & Ravand, 2016) investigated the differential performance of the UEE for different subpopulations of the same population, which is referred to as differential item functioning (DIF). To the best knowledge of the authors, few studies have explored the factor structure (invariance) of the UEE tests in general and construct validity of the PhD UEE in particular.

Factor Structure and Construct Validity

According to Messick (1995), although validity is a unitary concept, six distinct aspects of construct validity are highlighted as a means of addressing central issue implicit in the notion of validity as a unified concept. These six aspects of construct validity are *content*, *substantive*, *structural*, *generalizability*, *external*, and *consequential*. In the present study structural and generalizability aspects of validity are addressed through exploring the factor structure (invariance) of the PhD UEE. The generalizability aspect of validity concerns the principle of invariance which is claimed to be the essence of validity argument in the human sciences. Rasch (1960) described invariance as: "The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison, and it should also be independent of which other stimuli within the considered class were or might also have been compared" (p. 332). The group of persons who take a given test is a sample of the population of all possible test takers, and the items are a sample of all possible items which could be included in the test. The item and person invariance need due attention in generalizing the interpretation of the test scores.

The structural aspect of validity addresses the degree to which the scoring model matches the structure of the test (Messick, 1995). As Adam & Wu (2007, p.21) argued, “an aggregated item score is meaningful just when all the test items tap into the same latent variable. Otherwise, one outcome score for different dimensions is uninterpretable, since the same total score for students A and B could mean that student A scored high on latent variable X, and low on latent variable Y, and vice versa for student

B". Hence, the dimensionality of the test is a determining factor in the choice of its scoring model.

Exploring the factor structure of the test would help to match the scoring policy of the test with the nature of the construct and investigating invariance of the factor structure across subpopulations provides evidence on whether the results are an artifact of the sample studied or they are sample-independent.

The significance of the Study

The present study is significant in many aspects. First and foremost it explores the validity of a high stakes test by investigating its factor structure and invariance of the factor structure across different subpopulations. Studying the factor structure invariance of the PhD UEE can provide evidence on fairness and consequently, in Messick's (1989) terms, *generalizability* aspect of the construct validity of the test. Generalizability holds when a test measures the same construct across test takers belonging to different subpopulations of the same population (e.g., gender, ethnic background, language proficiency level). Factor structure study of the PhD UEE would also be of practical interest in that it would shed light on whether the score-reporting policy currently practiced by the MO matches the internal structure of the PhD UEE. In other words, the present study would provide evidence on the *structural aspect* of construct validity, as proposed by Messick (1989). Currently, a single percent-correct score is reported for the whole GE section of the test, which matches a unidimensional construct. However, if the results of the present study show that the construct measured by the GE section is multidimensional, then, to be congruent with the requirement of the structural aspect of construct validity, a single score should be reported for each section.

On a more general level, the present study would shed lights on the nature and dimensionality of language proficiency. There is no clear consensus about the definition, nature as well as the dimensionality of

language proficiency construct. According to Farhadi and Abbasian (2000) despite the efforts made to define language proficiency theoretically as well as operationally, there is no agreed-upon definition. They believe that some of these unclear definitions have triggered the development of unidimensional vs. multidimensional models of language ability.

The divisibility and dimensionality debates of language proficiency have revolved around the number of factors accounting for its underlying structure. More specifically these debates have concerned whether language proficiency is a unitary factor, a set of uncorrelated factors, or an overarching factor consisting of some correlated factors. The results of the present study hopefully would shed light on another equally important but still unresolved issue: factor differentiation. Factor differentiation refers to decrease or increase in the magnitude of factor correlations among different levels of language proficiency. Specifically, the debate is "whether or not the dimensions of language ability become more or less differentiated as a function of increasing examinee proficiency" (Shin, 2005, p. 31)

Review of the Literature

Divisibility of Language Proficiency

The divisibility of language proficiency is concerned with determining how many factors account for its underlying structure. The divisibility debate has concerned whether language proficiency can be better modeled with a unitary factor model, an extreme divisible model (a model with uncorrelated factors), or a partially divisible model (a model which consists of one general factor or *g* factor and some smaller factors [Barbour, 1983]). Oller (1978) conceptualized language proficiency as one general factor and argued for the unitary competence hypothesis (UCH). Oller (1979) conducted a study on 159 Iranian adult students in Tehran who took the TOEFL test. The result of his study approved the indivisibility hypothesis of language proficiency. Based on the similar performance of the test takers on a variety of language tests in different modalities, he made a strong case for the existence of a unitary

competence, which Oller dubbed *expectancy grammar*. Oller's UCH was criticized by Farhady (1983) and Carrol (1983) due to methodological flaws. According to these critics, Oller used principal component analysis in which error variance components were included in the analysis, and using this analysis might lead to overestimation of the first factor. They also called UCH into question because Oller used unrotated factor analysis and consequently took the first factor as a general factor. Some of the subsequent studies investigating the nature of second language proficiency have supported the multicomponent nature of language proficiency which consists of one general higher-order factor as well as several distinct first-order factors (Oller, 1983; Carrol, 1983, Bachman et al., 1990, Fouly et al., 1990). Still, some other studies have found correlated first-order factors (e.g., Bachman & Palmer, 1982; Sasaki, 1996; Shin, 2005) for the structure of L2 proficiency. As Vollmer (1983) pointed out, the multidimensional model of language proficiency consists of two versions: the strong and weak version. The robust and multidimensional version expected the existence of 16 skills for the language knowledge and the weak version assumed four components for the language proficiency. Furthermore, Zhan (2010) stated that most language teachers follow the traditional definition of language proficiency and are more familiar with this definition: "language proficiency comprises linguistic skills in the four core curricular areas: listening, speaking, reading, and writing" (p. 120).

Dimensionality of Language Proficiency

Another equally important but still unresolved issue about language proficiency is related to its factor differentiation: the decreasing or increasing order in the magnitude of factor correlations among different levels of language proficiency. Some of the factor structure studies which focused on the dimensionality of language proficiency tests found that the factor structure of the tests varied across test takers with different proficiency levels. Swinton and Powers (1980), conducting a factor analytic study of the Test of English as a Foreign Language (TOEFL),

concluded that candidates' language proficiency level and the degree of factor differentiation of the test are positively related. In other words, as candidates' proficiency level increases, the factor differentiation exhibited by the test increases as well, and vice versa.

Similarly, Ginther and Stevens (1995) conducted a series of multiple-group SEM analyses to investigate the construct validity of the Advanced Placement Spanish Language Examination. In line with Swinton and Powers, they found that lower levels of candidates' language proficiency led to lower factor differentiation and vice versa. However, other studies have found a negative relationship between factor differentiation and level of language proficiency (Hosley &Meredith, 1976; Kunnan, 1992; Farhadi & Abbassian, 2000; Romhild, 2008). There are still studies which have found that the structure of language proficiency remained the same across different proficiency levels (Shin, 2005; Stricker & Rock 2008). Related to both factor structure invariance and factor differentiation, Alderson (1991) stated that "language proficiency is both unitary and divisible at the same time" (p. 18). He believed that the nature of language proficiency depends on the level of language proficiency. He argued that language proficiency seems to be more unidimensional at higher levels, and more multifactorial at lower and intermediate levels.

Validity Studies on Language Proficiency Tests

Many research studies have been devoted to the investigation of the construct validity of high-stakes tests such as TOEFL (e.g. Farnsworth, 2013; Sawaki, Stricker, & Oranje, 2008; Stricker, Rock, & Lee, 2005), ECPE and MELAB (Jiao, 2004; Saito, 2003; Wagner, 2004, Wang, 2006), etc. Using Structural Equation Modeling (SEM), Shin (2005) studied the relationship between proficiency level and the structure of the Test of English as a Foreign Language (TOEFL) and the Speaking Proficiency in English Assessment Kit (SPEAK). She examined four models (a second-order factor model, a correlated-factor model, a single general factor model, and an entirely divisible model) and established a second-order factor model as the baseline model. The result of her study showed that the

structures of TOEFL and SPEAK were invariant across different proficiency groups and supported none of the hypotheses related to the dimensionality of language proficiency. Stricker and Rock (2008) studied the factor structure invariance of TOEFL across test takers with different native languages and different amounts of exposure to English. They employed confirmatory factor analysis using SEM for data analysis and postulated four models: a single factor model, a two-factor model, a four-factor model, and four first-order and a higher-order factor models. In line with Shin's findings, they found that the test was invariant across different subgroups of test takers.

Wang (2006) investigated the factor structure invariance of Certificate of Proficiency in English (ECPE), and the Michigan English Language Assessment Battery (MELAB) tests across different genders. SEM analysis was used in order to conduct a multigroup analysis. In this study, a unitary factor model was examined for both tests. The results indicated both tests were equivalent across males and females and it was evidence for the fairness of these two tests. In'nami and Koizumi (2011) conducted an SEM study on the factor structure of the listening and reading comprehension sections of the Test of English for International Communication (TOEIC). They tested a higher-order, a correlated, an uncorrelated, and a unitary factor model. The results supported the correlated factor model which in turn supports the divisibility of language proficiency.

Furthermore, the results of the multigroup analysis suggested the invariance of the correlated model across different samples. However, Innami et al. (2016) investigated the Test of English for Academic Purposes (TEAP) and compared it with the TOEFL test. Using confirmatory factor analysis, they tested four models (unitary, correlated, receptive-productive, and higher-order factor model) and found that the higher-order factor model shows the best fit model. The results of their study indicated that there is a close relationship between TEAP and

TOEFL tests and it was evidence for construct validity of this high stake test.

Validity Studies on UEE Tests

Barati and Ahmadi (2010) investigated gender and significant differential item functioning (DIF) on the bachelor's UEE for the applicants into English programs. The study utilized a one-parameter IRT model with a sample of 36000 test takers who sat the test in 2004. The findings of their study confirmed the presence of DIF in some of the items of this high-stakes test. Similarly, using the Rasch model, Ravand and Firoozi (2016) investigated the construct validity of the 2009 version of the Master's UEE for the applicants into English programs. They found that the test as a whole did not show unidimensionality. As a result, they decided to analyze different sections of the test namely reading, grammar, and vocabulary separately. According to authors, lack of the invariance of the person measures was another piece of evidence against construct validity of the test.

However, to the best knowledge of the authors, there have been very few validation studies on the PhD UEE (e.g., Ahmadi et al., 2015; Alibakhshi & Ghandali, 2011). Ahmadi et al. (2015) conducted a concurrent triangulation mixed method research study to check the reliability and validity of the PhD UEE based on Kane's (1992) argument model and Bennett's (2010) theory of action. The result of their study indicated that validity and reliability of this high-stakes test were under the question regarding the test takers' dissatisfaction with test administration conditions including test venue, testing time and difficulty level of the IPEET items. Moreover, the results of Logistic Regression (LR) showed 12 items of this high stake test were flagged for DIF.

Regarding the paucity of the studies exploring construct validity of the PhD UEE in particular and the debates on the divisibility of language proficiency and also inconsistencies in previous findings on the relationship between degree of language proficiency and language test functioning, in general, this study intends to investigate the factor structure

of the PhD UEE and its (in)variance across different proficiency levels using SEM. Several features of SEM give it an edge over conventional statistical methods such as correlation and regression: (a) capability of either assessing or correcting for measurement error, b) incorporating both observed and latent variables, c) modeling and estimating multivariate direct and indirect relations.

For the present study the following research questions were posed:

1) What is the factor structure of the Ph.D. UEE?
Factor structure studies on high-stakes proficiency tests have mostly compared the fit of a series of factor models such as a unitary factor model, a correlated factor model, and a higher order factor model (e.g., Innami & Koizumi, 2011; Romhild, 2008; Sawaki et al., 2009). To keep continuity with these studies in the present study, a fit of four models were compared: a correlated three-factor model a unitary, an uncorrelated, and a higher-order factor model.

2) Is the factor structure of the Ph.D. UEE invariant across high and low proficiency groups?
The answer to this question can have implications for the relationship between test takers' proficiency level and degree of factor differentiation.

Data

This study is based on the data from 1009 test takers (573females and 436 males) who took the Ph.D. UEE to seek admission into English programs at state universities in Iran in March 2013. The data provided by MO were item-level data. PhD UEE applicants take three sets of questions: general English (GE), scholastic aptitude, and content knowledge questions.

The participants mostly aged between 22 and 35 and they had their Master's education mostly in four university types in Iran: (a) state universities which usually do not charge any tuition fees, (b) Azad universities which charge tuition fees (c) Non-profit Non-government

universities which charge tuition fees, much lower than those of Azad universities, and (d) Payam-e-Noor universities which charge tuition fees as much as those of Non-profit Non-government universities but do not offer regular classes.

The data for the present study were from the GE section of the PhD UEE. There were 70 GE items including ten grammar, 20 vocabularies, ten cloze, and 30 reading comprehension items. All the questions were multiple choice, and the test takers had to complete the GE section in 105 minutes.

Analysis Procedures

To come up with the factor structure of the PhD UEE, first, the data were subjected to exploratory factor analysis (EFA) in a series of steps: *eigenvalue* criterion, scree plot, and *simple structure* criteria were used to determine the number of factors underlying the test. According to the EFA results, confirmatory factor analysis (CFA) was run in a series of steps to come up with the appropriate factor structure of the test. Then four hypothesized models were tested and compared to determine the best baseline model: (a) a unitary factor model (Figure 1), (b) an uncorrelated three-factor model (Figure 2), (c) a correlated three-factor model (Figure 3), and (d) a higher-ordered factor model (Figure 4).

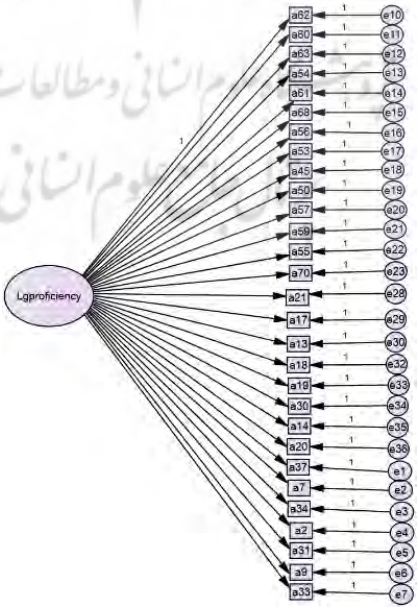

Figure 1. Unitary Model
Note: lgprof, is., language proficiency

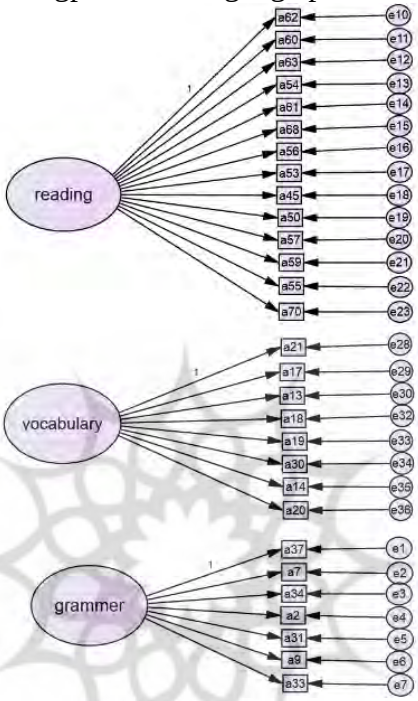

Figure 2. Uncorrelated Three-Factor Model

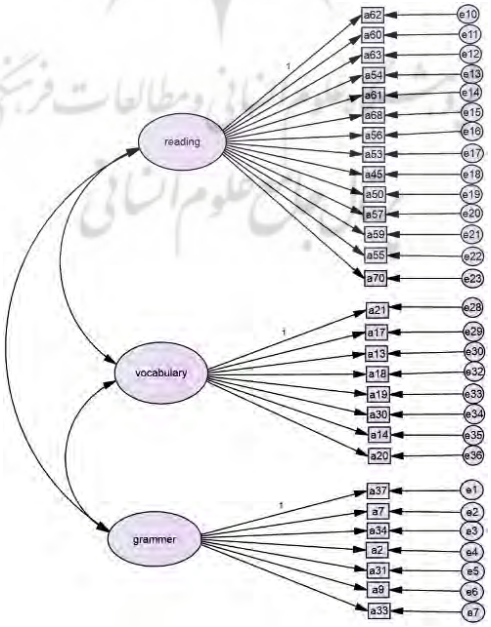

Figure 3. Correlated Three-Factor Model

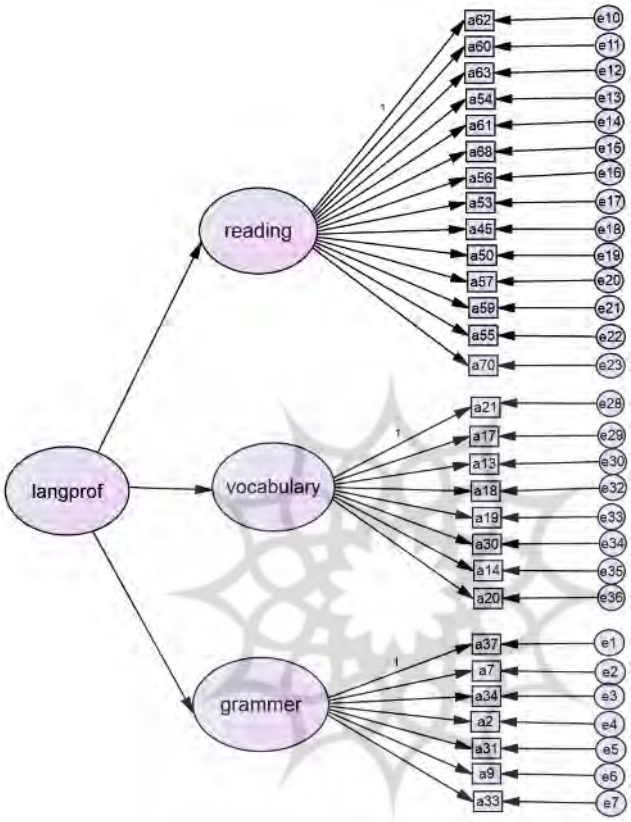

Figure 4. Higher-Ordered Factor Model
Note: langprof, is., language proficiency

Furthermore, to examine the factor structure invariance of the test across two proficiency groups, a multigroup CFA was conducted on the baseline model obtained in the previous stage of analysis. For multi-group CFA, as required by the second research question, the sample was split into three groups based on the examinees' grade point averages (GPA) at Bachelor's level. To increase the contrast between the groups, the mid group was removed from further analysis.

Each formulated CFA model was evaluated based on some conventional criteria which are used widely in the literature: the ratio of chi-square (x^2) to degree of freedom (*df*),x^2/df , where values of 2.0 and

below are recommended as indication of good model fit (Byrne, 1989; Ullman, 2001), Comparative Fit Index (CFI), Tucker-Lewis index (TLI), and Root Mean Square Error of Approximation (RMSEA). The reasonable minimum cutoff value for CFI should be 0.90 (Bentler, 1992; Hoyle, 1995). Hu and Bentler (1999) advised the revised cut-off value of 0.95 for this index. However, some researchers believe that the cutoff values suggested by Hu and Bentler's are too rigorous (Beauducel & Wittman, 2005; Yuan, 2005). Moreover, the TLI cutoff value ≥ 0.95 indicates a well-fitting model (Hu& Bentler, 1999). RMSEA index suggested by Steiger and Lind (1980) is regarded as informative criteria in covariance structure modeling (Byrne, 2013). The proposed values of 0.05 or below and 0.08 or below are considered as close fit and adequate fit, respectively (Browne & Cudeck, 1993). SPSS 22 and AMOS 20 were used to conduct the analyses.

Results

Exploratory Factor Analysis

As a preliminary step for CFA, EFA was performed to determine the number of factors needed to explain the relationship among the observed variables (the items of the test). First of all the eigenvalues were checked. In the second step, the scree plot was consulted. Then, *the parallel analysis* was run using Monte Carlo simulation. The final decision on the number of factors to retain was made based on interpretability and *simple structure*. According to the Kaiser Criterion (K1 rule), proposed by Kaiser (1960), all factors with an eigenvalue greater than 1.0 or more were retained. This method has been criticised since it leads to over-estimation of the number of factors (Fabrigar et al., 1999; Zwick & Velicer, 1982, 1986). In the current study, based on the K1 rule, there were about 27 factors which explained the clustering of the data. However, since the extraction of this many factors was not compatible with the current understanding of language proficiency, Scree Plots of the eigenvalues were consulted. The plot suggested that three factors accounted for the variance in the item-

level data. Horn's (1965) parallel analysis (PA) can be considered as an alternative to K1 rule (Garrido, Abed, & Ponsoda, 2012; Velicer, Eaton, & Fava, 2000). Based on this method some data were randomly generated, comparing this data with the size of eigenvalues leads the researchers to retention of only those eigenvalues which exceed the randomly generated data values. The results of parallel analysis suggested retention of 10 factors.

According to the simple structure and current understanding of language proficiency, the results from three, four, and five-factor solutions were compared. Overall, the three-factor *oblique* solution yielded the simple structure and the most interpretable results. The three factors were named based on the items which loaded onto them. The first factor on which 12 reading items loaded was interpreted as the Reading factor, the second factor on which eight vocabulary items loaded was interpreted as the Vocabulary factor, and the third factor on which seven grammar items had high loadings was interpreted as the Grammar factor.

Confirmatory Factor Analysis

Establishing Baseline Model

As a preliminary step for testing factorial invariance, four hypothesized models were formulated and compared in order to select the baseline model.

1) A unitary model (Model A, Figure 1): This model consists of only one factor, language proficiency, in which all the 29 items are affected by just one single factor. This is equivalent to saying that the three factors underlying language proficiency- Reading, Vocabulary, and Grammar- are indivisible.
2) An uncorrelated three-factor model (Model B, see Figure 2): In this model the three factors are independent of and entirely unrelated to each other.
3) A correlated three-factor model (Model C, see Figure 3): In this model, it is hypothesized that three factors underlying performance on PhD UEE are correlated.

4) A higher-order factor model (Model D, see Figure 4): This model is identical to the correlated three-factor model, except for the relationship between the three factors which is modeled through a second-order general factor. This model hypothesizes that language proficiency affects test takers performance on all the subsets including Grammar, Vocabulary, and Reading. It consists of a single higher-order factor and three first-order factors.

Formulation of the higher-order model was motivated by the literature and high correlations between the factors according to the EFA results. EFA showed that correlations between Grammar and Vocabulary, Grammar and Reading, and Reading and Vocabulary were 0.52, 0.56, and 0.62, respectively. Statistically, a higher-order model is more parsimonious than a correlated factor model and is to be preferred. However, with only three first-order factors (as is the case in the present study), the model is *just identified* and therefore not distinguishable from a three-factor correlated model (Rindskopf & Rose, 1988). Therefore, the fit indices for the two models are expected to be precisely the same.

The four models were compared based on the criteria above. Table 1 shows the fit indices for the four models estimated for low and high groups, separately.

Table 1.

Fit Indices for Four Models

	Low					High				
Fit Indices	x^2	CFI	TLI	x^2 /df	RMSEA	x^2	CFI	TLI	x^2 /df	RMSEA
Model A	345.14	.862	.836	1.255	.022	396.54	.840	.838	1.442	.030
Model B	394.94	.764	.761	1.436	.029	398.60	.837	.835	1.449	.030
Model C	285.90	.973	.969	1.051	.010	280.95	.988	.984	1.033	.008
Model D	285.90	.973	.969	1.051	.010	280.95	.988	.984	1.033	.008

As can be seen in Table 1, the RMSEA values for all models in both groups are below 0.05 indicating good fit (Brown & Cudeck, 1993). The x^2/df ratios are well below the cut-off point of 2.0 indicating a nonsignificant chi-square. The CFI and TLI index values for Model A and Model B are below the cut-off point of .95 which are not satisfactory and indicate poor model fit across both proficiency levels. Therefore, these two models will not be considered further. Furthermore, as expected, the fit indices for Model C and Model D are the same and within the excellent-fit range. Concerning both parsimony and substantive meaningfulness (Byrne, 1994) the higher-order model should be selected as the baseline model. A higher-order model is preferable to a correlated factor model in that covariance among the first-order factors is explained by a higher-order factor. As alluded to before, a higher-order model is also more consistent with the extant literature on the structure of language proficiency (e.g., Bachman & Palmer, 1981a; Sawaki, Stricker, & Oranje, 2009; Shin, 2005). However, since the higher-order factor model had only three first-order factors and the model was just-identified, its fit could not be tested (In'nami & Koizumi, 2011). Therefore, the correlated three-factor model was selected as the baseline model.

Test of Factorial Structure Invariance Across the Two Proficiency Levels

In seeking evidence of factor structure invariance for the PhD UEE across the two proficiency levels, multigroup confirmatory factor analysis was conducted. The process of checking for measurement and structural invariance involves examining a series of increasingly restricted models. "The measurement issues concerns the invariance of cross-group factor loadings and the error variances, while the structural issues address theinvariance of factor variances and covariances across groups" (Bae &Bachman, 1998, p. 385). The steps involve checking: (a) configural invariance: Configural model invariance requires that the same number of factors are represented and each common factor is associated with identical item sets across the groups, (b) invariance of factor loadings, (c) invariance

of factor loadings, and the error variances, (d) invariance of factor loadings, the error variances, and factor variances, and (e) invariance of factor loadings, the error variances, factor variances, and factor covariances.

First, the configural model, as a preliminary step in checking *measurement invariance* (Horn, McArdle, & Mason, 1983), was estimated for the two groups simultaneously without imposing any equality constraints on the parameters. In this step, the factor loadings were freely estimated, and no parameter constraints were specified. However, the relationships between observed variables and factors were set identical across the two groups. As seen in Table 2, all fit indices for Model 1 (e.g. CFI= .96, RMSEA=.008, TLI= .96) indicate a good overall fit. Based on this information we can approve the multigroup correlated model of UEE as the configural model.

Second, the analysis of measurement invariance was conducted because the configural invariance does not provide sufficient evidence as to whether the test items measure the same construct across different groups. Therefore, we proceeded with checking measurement invariance by constraining all factor loadings to be equal across the two groups. As one can observe from Table 2 the CFI, RMSEA, and TLI values for this model are minimally different from the configural model, suggesting that all factor loadings were invariant across the two proficiency levels.

Third, both factor loadings and measurement error variances were constrained to be equal across the two groups. This step is more stringent than the previous step wherein only the first-order factor loadings were constrained. The fit indices in Table 2 show that this model fits the data very well (CFI= .960, RMSEA= .008, TLI=.957) proposing that both factor loadings and error variances be invariant across the two proficiency levels.

Fourth, in order to check structural invariance, additional constraints were added, and factor variances were constrained across the two groups. This model as shown in Table 3 fit the data well (CFI= .958,

RMSEA= .008, TLI=.955), suggesting that all factor loadings, error variances, and factor variances were invariant across the two proficiency levels. “When this level of invariance holds, all group differences on the items are due only to group differences on the common factors” (Chen, Sousa, & West, 2005; p .474)

Finally, the most stringent constraint was added, and the factor covariances of the Model 5 were constrained to be equal across the samples. This model shown in Table 2 fit the data well (CFI= .963, RMSEA= .008, TLI=.959) suggesting that all factor loadings, error variances, factor variances, and factor covariances were of equal size across the two proficiency levels.

Table 2.

Fit indices for Correlated Model for Cross-Validation

	df	x^2	TLI	CFI	RMSEA
Model 1. Baseline model	748	790.375	.964	.969	.008
Model 2. Factor loadings equal	774	816.326	.965	.969	.007
Model 3. Factor loadings and error variance equal	803	857.781	.957	.960	.008
Model 4. Factor loadings, error variance, and factor variance equal	806	863.292	.955	.958	.008
Model 5. Factor loadings, error variance, factor variance, and factor covariance equal	783	834.233	.959	.963	.008

Furthermore, Models 1 to 5 was compared using chi-square difference tests and CFI to check whether the correlated three-factor model was invariant across the two groups (see Table 3). As Byrne (2013) stated, one of the main steps in checking factor invariance is computing the $x2$ difference and ΔCFI tests. The chi-squares differences which are significantly different and CFI values which have difference above .01 show a significant difference between the models. According to Cheung and Rensvoled (2002) CFI values ≤ .01 indicate that the invariance hypothesis should not be rejected.

Table 3.
Chi-Square and CFI Difference Tests

Model comparison	Δx^2	Δdf	Sig.	Δ CFI
Model 1 vs. Model 2	25.951	26	.09	.000
Model 1 vs. Model 3	67.406	55	.05	.009
Model 1 vs. Model 4	72.917	58	.10	.011
Model 1 vs. Model 5	43.858	35	.05	.006

According to the criteria mentioned above the chi-square values show nonsignificant differences, and differences between CFI values of the constrained models and configural model are well below the 0.1. These results suggest that imposing a series of increasingly restrictive constraints on the factor loadings, error variance, factor variance, and factor covariance across two groups of low and high proficiency level do not lead to significant drop in fit of the model. For example, for Model 2, both the $x2$ difference test and CFI difference test claim for evidence of invariance (See Table 3).

Discussion

This study examined the factor structure of the UEE for PhD applicants into the English programs in state universities in Iran. This study aimed to examine specifically two hypotheses, one related to the dimensionality of language proficiency and the other concerned the relationship between the structure of language proficiency and test takers' level of language proficiency (low and high-level groups). In order to answer the first research question, four models were developed. Each of these models was tested separately and compared with other models. The correlated three-factor model was selected as the baseline model for both proficiency groups for two reasons. First, the correlated three-factor and higher-order factor models fit the data for both samples better than the unitary and uncorrelated factor models.

Moreover, the findings of some previous studies (e.g., Backman & Palmer, 1982; Bae & Backman, 1998, In'nami and Koizumi, 2011, Shin,

2005) have confirmed that the correlated factor models are not significantly different from higher-order factor models. Second, the higher-order model had an identification problem because it had only three first-order factors. The results of the present study indicated the general English proficiency of the Ph.D. UEE candidates is measured through three components of vocabulary knowledge, grammar knowledge, and reading comprehension which is evidence against the UCH model of language proficiency suggested by Oller (1979). The finding of this study supports the multidimensionality view of language proficiency which is in line with Song (2008) who also found evidence in support of the multidimensionality of language proficiency. The multicompenentional view of language proficiency supported in the present study is also in line with Zhang (2010). Zhang also rejected the UCH model of language proficiency and considered a four skill model composed of listening, speaking, reading, and writing underlying language proficiency. Also, In'nami and Koizumi (2011) compared the unitary, uncorrelated, correlated, and higher order model and finally coming up with the correlated model as the final model. They believed that the unitary and uncorrelated factor models were statistically less favorite than correlated and higher order model.

The fit of the correlated three-factor model has got implication for the validity of the PhD UEE. According to Messick (1989), the structural aspect of construct validity requires that score reporting policy of any given test should match the structure of the test. As the results of the present study supported a three-factor model (Reading, Vocabulary, Grammar), three different scores should be reported. However, the measurement organization (MO) reports a single score for the GE section of this high stakes test. Consequently, the structural aspect of the construct validity of the test is under question.

To test for the factor structure invariance of the PhD UEE across the two proficiency levels, the correlated three-factor model was postulated as the baseline model in both groups, and then some gradually increasing constraints were imposed on the model. The results of checking

measurement and structural invariance, based on Meredith's (1993) classification of different levels of factorial invariance, showed that the structure of the PhD UEE was *strictly* invariant across the two groups. Specifically, the results showed that three factors of Reading, Vocabulary, and Grammar held for the two language proficiency groups (low and high). Also, as seen in Table 3, the results of the multigroup analysis showed invariance in factor loadings, error variances, factor variances, and factor covariances for the correlated three-factor model across the two groups. The study results suggest that the test tasks of the PhD UEE performed equally for both low and high-level proficiency groups. The results suggest that group members did not have a differential influence on the structure of the test. Since there are no factor structure studies on the PhD UEE, the results of this study cannot be compared with the related literature. However, the results have implications for the factor structure invariance of the test and challenge both the increasing and decreasing factor differentiation hypotheses. In the current study, the finding of *strict* measurement invariance suggested that group members not have a differential influence on the structure of the test.

Consequently, neither the increasing factor differentiation nor the decreasing factor differentiation hypothesis was approved. In other words, three factors of Reading, Vocabulary, and Grammar, underlay the performance on the PhD UEE and these three factors held for the two language proficiency groups (i.e., low and high). This finding confirms those of Shin (2005) and Stricker and Rock (2008) who argued that the structure of tests is the same across different proficiency groups. However, the results of this study are not in line with those of Ginther and Stevens (1995) and Kunnan (1992). Ginther and Stevens (1995) conducted a factor structure study on the Advanced Placement Spanish Language Examination and found the most significant differentiation in factor structure for the high proficiency group and the lowest degree of factor differentiation for the low-level proficiency group. On the other hand,

Kunan (1992) found evidence for a negative relationship between the level of language proficiency and degree of factor differentiation.

Factor structure invariance of the PhD UEE has implications for generalizability aspect of the construct validity as well as the fairness of this high stakes test. The Messick's (1989) generalizability aspect of construct validity requires the test to measure the same construct across different subpopulations. The study results suggest that the test tasks of the PhD UEE performed equally for both low and high-level proficiency groups. In other words, the PhD UEE fairly measures the same construct across different subpopulations, and the test takers' performance is comparable.

Limitations and Suggestions for Further Studies

There are some limitations to the present study that are worth mentioning because they provide suggestions for further research. The GE section of the PhD UEE included grammar, vocabulary, and reading comprehension. However, most of the factor structure (invariance) studies on high-stakes tests in the literature have been composed of speaking, listening, reading, and writing. Therefore, comparing the results of the present study with those of the literature and making claims about componentiality of language proficiency and its factor differentiation based on these results should be carried out with caution.

In the current study, invariance was tested across two proficiency levels. To this purpose, the subjects were divided into two proficiency groups (low and high) only based on their Bachelors' GPAs. Future studies can use other grouping criteria such as university status, the candidates' gender, the field of study, and type of university they graduated from. Strong arguments for generalizability aspect of the construct validity of the test under study in the present explorations should be put off to the time when invariance based on gender, major, and type of university where the test takers received their Bachelor's education is tested.

It is notable that since state universities in Iran usually offer higher quality education and facilities and charge no tuition fees, there is a tight

competition to obtain a seat in these universities. Usually, applicants with better scores opt for these universities and only when they fail to secure a seat at state universities might they decide to continue their education at other types of universities. It is commonly believed that students who are educated at state universities are more knowledgeable due to the better quality of education they receive and the fact that studentship at state universities is the first job for most students, but most of those studying at the other three types of universities primarily hold other jobs and are seeking university degrees to get promotion at work. It would be interesting to check factor structure (invariance) of the test across groups with different types of education at Master's level, i.e., those who have graduated from State, Azad, non-profit non-government, and Payam Noor Universities.

References

Ahmadi, A., Darabi Bazvand, A., Sahragard, R., & Razmjoo, A. (2015). Investigating the Validity of PhD. Entrance Exam of ELT in Iran in Light of Argument-Based Validity and Theory of Action. *Journal of Teaching Language Skills*, *34*(2), 1-37.

Alavi, T. (2012). The Predictive Validity of Final English Exams as a Measure of Success in Iranian National University Entrance English Exam. *Journal of Language Teaching and Research*, *3*(1), 224-228.

Alderson, J. C. (1991): Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (p. 18). Singapore: SEAMEO Regional Language Centre.

Alibakhshi, G., & Ali, H. G. (2011). External Validity of TOEFL Section of Doctoral Entrance Examination in Iran: A Mixed Design Study. *Theory and Practice in Language Studies*, *1*(10), 1304-1310.

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

Bachman. L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L. F. &Palmer. A. S. (1981a). The construct validation of the FSI oral interview. *Language Learning, 31*(1), 67-86.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(1), 449–465.

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, England: Cambridge University Press.

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

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the Subject Area: A Study of the Iranian National University Entrance Exam. *The Journal of Teaching Language Skills (JTLS), 2*(3), 1-22.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high-stakes tests: the effect of field of study. *International Journal of American Linguistics (IJAL), 9* (3), 27-49.

Barbour, R. P. (1983). *An exploratory study of the hypothesis of divisible versus unitary competence in second language proficiency* (Doctoral dissertation, University of British Columbia), 1-147.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with a slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41-75.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8* (1), 70-91.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*(3), 400.

Birjandi, P., & Amini, M. (2007). Differential item functioning (test bias) analysis paradigm across manifest and latent examinee groups (on

the construct validity of IELTS). *Journal of Human Sciences*, *8*(2), 1-20.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B.M. (1989): *A primer of LISREL: basic applications and programming for confirmatory factor analytic models*. New York: Springer Verlag.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. University of Ottawa, Canada. Sage.

Byrne, B. M. (2013). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.

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

Carroll, J. B. (1983). Psychometric theory and language testing. *Issues in Language Testing Research*, 80-107.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*(3), p.474.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4(*3), 272.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. *Issues in Language Testing Research*, 11-28.

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

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274-291.

Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning, 40*(1), 1-21.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2012). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, in press. Epub ahead of print retrieved December 10, 2012.

Ginther, A., & Stevens, J. (1995). Language Background, Ethnicity, and the Internal Construct Validity of the Advanced Placement Spanish Language Examination. *Education Resource Information Center (ERIC).* 1-27.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30(*2), 179-185.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance? *Southern Psychologist, 4*(2), 179-188.

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Sage Publications.

Hu, L-T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling, 6*(1), 1–55.

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

Jamalifar, G., Tabrizi, H. H., & Chalak, A. (2014). Islamic Azad University Entrance Examination of Master Program in. *The Iranian EFL Journal, 29*(1), 386.

Jiao, H. (2004). Evaluating the dimensionality of the Michigan English language assessment battery. *Spaan Fellow Working Papers in Second or Foreign Language Assessment. 2004*(2), 27-155.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement, 20*(1), 141-151.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kline, R. B. (1989). *Principles and practice of structural equation modeling*. London: The Guilford Press.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30-49.

Mahmoudi, L., & Bakar, K. A. (2013). Iranian Pre-university English Teachers' Perceptions and Attitudes towards the Iranian National University Entrance Exam: A Washback Study. *International Journal of Education & Literacy Studies, 1*(2), 47.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*(4), 525-543.

Mesick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

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. (1979). *Language tests at school: A pragmatic approach*. Addison-Wesley Longman Ltd.

Oller, J. W., & 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.

Ravand, H., & Firoozi, T. (2016). Investigating Validity of UEE using the Rasch Model. *International Journal of Language Testing, 6*(1), 1-23.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51-67.

Römhild, A. (2008). Investigating the invariance of the ECPE factor structure across different proficiency levels. *Spaan Fellow*, 6(1), 29-54.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2003*(1), 39-82.

Salehi, H., & Yunus, M. M. (2012). The washback effect of the Iranian universities entrance exam: Teachers' insights. *GEMA: Online Journal of Language Studies, 12(2), 609-628.*

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

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

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing,* 26 (1), 5-30.

Song, M. Y. (2008). Do distinct subskills exist in second language (L2) comprehension? A structural equation modeling approach. Language Testing, 25(4), 435-464.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. In annual meeting of the Psychometric Society, Iowa City, IA. 758(1), 424-453.

Stricker, L. J., & Rock, D. A. (2008). Factor Structure of the TOEFL Internet-Based Test across Subgroups. *ETS Research Report Series*, *2008*(2), i-38.

Stricker, L. J., Rock, D. A., & Lee, Y. W. (2005). Factor structure of the languedge™ test across language groups. *ETS Research Report Series*, *2005*(1), i-43.

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. *ETS Research Report Series, 1980*(2), i-79.

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell, *Using Multivariate statistics* (pp. 653–771). Needham Heights, MA: Allyn and Bacon.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A re-review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and solutions in human assessment* (pp. 41-71). Springer US.

Vollmer, H. J. (1983). The structure of foreign language competence. *Current developments in language testing*, 3-30.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment 2004*(1), 1-155.

Wang, S. (2006). Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender. *SPAAN FELLOW, 4* (1), 41-56.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115-148.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), p.120.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432.