Journal of Modern Research in English Language Studies
Print ISSN: 2676-5357
Online ISSN: 2676-5985

# English Proficiency Homogenization Using Norm-referenced and Criterion-referenced Tests: A Structural Equation Modeling Approach

Kazem Barzegar[1], Amir Reza Nemat Tabrizi [2*], Manoochehr Jafari Gohar[3], Fereidoon Vahdany[4]

[1] Ph.D. Candidate at Payame Noor University, Department of Linguistics and Foreign Languages, Tehran, Iran, *kbarprof@ssu.ac.ir*
[2*] Ph.D. in TEFL, Payame Noor University, Department of Linguistics and Foreign Languages, Tehran, Iran, *arnemati@yahoo.com*
[3] Ph.D. in TEFL, Payame Noor University, Department of Linguistics and Foreign Languages, Tehran, Iran, *jafarigohar2007@yahoo.com*
[4] Ph.D. in TEFL, Payame Noor University, Department of Linguistics and Foreign Languages, Tehran, Iran, *frvahdany@yahoo.com*

## Abstract

This correlational research used the association between norm-referenced and criterion-referenced tests to predict CRT scores on the basis of NRT scores, homogenize English proficiency of university students, and design a structural equation modeling approach between NRTs and CRTs. The participants were 210 allied health EGP (English for General Purposes) students who were assigned to three levels of language proficiency using Cambridge Placement Test. Researcher-made midterm and final exams, focusing on grammar, vocabulary, and reading comprehension were conducted. Results showed significant positive correlations between the NRT and CRTs. Structural Equation Modeling (SEM) analysis indicated significant paths from NRT as the dependent latent variable to CRTs as independent latent variables. The scores on the components of the three latent variables including vocabulary, grammar, and reading, within three assessments (placement, midterm, and final) were considered as observed variables. Significant paths between NRT and CRTs suggested that complex interrelations between components of NRT and CRTs can be used to homogenize university students' proficiency in academic English courses using NTR scores to overcome problems related to individual differences. Consequently, in academic English courses, groupings based on university students' language ability using NRT scores would override groupings solely based on students' academic majors.

***Keywords***: Criterion-referenced Tests, Individual Differences, Language Proficiency, Norm-referenced Tests

Received 04 February 2020 Accepted 01June 2020
Available Online 15 November 2020 DOI: 10.30479/jmrels.2020.12495.1547

***Article Type: Research***
***Vol. 8, No. 1, 2021, 55-74***

## 1. Introduction

Language learners around the globe are grouped in linguistically homogenous English classes in language institutes using various placement tests available. However, university students are placed in Basic English, General English, and EAP courses in linguistically heterogeneous classes where students of different language proficiency levels are taught English by the same teacher using the same teaching source, teaching method, and educational technology. For example, Iranian students of different medical and allied health majors with different language abilities are placed in a Basic English class and taught by the same instructor using the same source and an identical pedagogic methodology. In such classes, according to Barzegar and Askari (2015), while some students can use advanced idiomatic English in a native-like manner, beginners may not know the ABCs of English communication. This has led to the failure of academic English teaching at different levels in the Iranian academia as suggested by Jamshidi Avanaki and Sadeghi (2013) who explored ELT in Iranian universities and the pertinent problems related chiefly to untrained instructors, classrooms, textbooks and also the instructional approaches that are commonly mastered at the theoretical rather than the practical level. If these students are grouped in Basic English and EGP courses using their scores on a placement test (NRT), their scores on their academic English courses may be both predictable and improvable. Norm-referenced tests (NRTs) are employed to assess a test-taker’s performance against that of other testees. Nonetheless, criterion-referenced tests (CRTs) are different from NRTs in that each testee’s performance is judged based on a pre-set collection of criteria or a standard (Lok, McNaught, & Young, 2015). The basic goal of these kinds of tests, the results of which are used in making decision about receiving a kind of certification or benefiting from pedagogical services, is to determine whether the candidate has mastered a certain skill or set of skills (Bond, 1996). Given that the features of some educational contexts impede the administration of one of these types of tests, and thus the administration of the other is favored in those situations, invaluable information will be offered by the results of one of the tests if it is proved that prediction can be made in one type of test scores considering performance on another. When students of various different English proficiency levels are exposed to different teaching protocols, the final fulfillment of the groups is more rational.

The present study explored whether a CRT score can be predicted based on the results of an NRT one using “Structural Equation Modeling” (SEM). According to Winke (2014), the application of SEM as an important research tool for examining the similarity and strength of theories and hypotheses in applied linguistics and second language acquisition dates back to more than 15 years. SEM is an effective tool in assessing the impacts of various instructional modes on learning and investigating fairness in language testing and the influences of bilingual development. SEM has also been proved useful in assessing the influences of knowledge constructions such as metalinguistic knowledge that influence L2 performance (Isemonger, 2007). The application of SEM in analyzing second language acquisition (SLA) data increased during the last five years using various software programmes including PLS (*Ringle, Wende, & Becker,* 2015), Amos (Arbuckle, 2003), Mx, LISREL (J¨oreskog & S¨orbom, 1986), EQS (Bentler, 1990), Mplus (Muth´en & Muth´en, 2004), CALIS, and SEPATH.

This study aimed at testing the following three hypotheses:

1. There exists a statistically significant association between NRTs and CRTs as latent variables and their components as observed variables.
2. University students' CRTs scores can be significantly predicted based on their NRTs scores.
3. A structural model can be designed to significantly explain the relationship between NRTs and CRTs.

## 2. Literature Review

### 2.1. English Proficiency Homogenization

Academic English proficiency has been the topic of a number of studies round the globe regarding the correlation between language proficiency and other learner factors. The study by Barzegar et al. (2020) investigated the impact of English proficiency homogenization on linguistic proficiency of Iranian students of health as a channel of health promotion by reading English media. Also, Baker Smemoe and Haslam (2013) explored the impact induced by aptitude for language learning, learning context, and strategy use on second language pronunciation proficiency. Moreover, Iwashita et al. (2008) investigated the evaluated levels of proficiency in second language speaking to observe how distinct they are. In addition, Galaczi (2014) studied the interactional competence across various proficiency levels. Tomiyama (2009) also studied the effect of proficiency and age on second language attrition by the use of data from two siblings. Ortega (2003) explored the assessments of grammatical complexity and their relationship to L2 proficiency. Al-Gahtani and Roever (2012) contemplated on proficiency and consecutive organization of second language requests. Finally, Wen and Johnson (1997) expunged upon the correlation between Second Language learner variables and English proficiency. All these and similar studies try to show that differences in English proficiency levels of students engaged in academia cannot be ignored as they impact their manner and amount of EAP learning.

## 2.2. Application of SEM in Applied Linguistics

A review of literature showed that no scholar has so far embarked on academic English proficiency homogenization using "structural equation modeling" of the correlation between NRTs and CRTs. Yet, the study by Barzegar et al. (2020) explored the effect of linguistic homogenization on English proficiency of students of health in promoting health education. Hence, this was the first endeavor to investigate this correlation using SEM. Nevertheless, scholars have used SEM to examine other aspects of applied linguistics. According to Kline (2011), over the last 15 years, many studies have presented basic applications of SEM in applied linguistics: CFA (confirmatory factor analysis), and "path analysis", or "full latent trait model" testing. Some scholars like Kieffer (2011) and Mancilla-Martinez and Lesaux (2010) have presented latent growth curve modeling which is one of the many advanced uses of SEM. Winke (2014) also wrote a meta-analysis entitled: "Testing Hypotheses about Language Learning Using Structural Equation Modeling" in which they analyzed about forty articles on the application of SEM in applied linguistics. Zhang (2012) worked on linguistic knowledge structures using SEM concentrating on the way knowledge structures, as independent predictor variables, influence the language acquisition process. These structures dealt with working memory, metacognitive knowledge, awareness for phonology, or grammatical knowledge, and the dependent variables were assessments of reading comprehension, understanding discourse, listening comprehension, reading fluency, writing proficiency, or vocabulary development as examples. Xie and Andrews (2013) expunged upon areas of language testing using college or university students of English as participants, using convenient sampling. In these studies, the researchers investigated the factors that influence test preparation. Yang (2012) elucidated the impacts of test-taking strategies on improving performance on writing test. Song (2012) examined the influence of note-taking on performance on listening test. Additionally, Aryadoust (2010), Phakiti (2008a), Phakiti (2008b), and Song (2008) focused on construct-validation. Moreover, Aryadoust (2010), Phakiti (2008a) and Phakiti (2008b) used CFA to elucidate the subdivisions of larger constructs. Some other researchers have focused on willingness to communicate (WTC) using SEM. For instance, Aidinlou and Ghobadi (2012) explored how WTC influences language development, but the researchers in the paper did not give any information on how they evaluated the language. Furthermore, Gallagher (2013) mentioned clearly defined dependent variables: perceived stress, while Fushino (2010) surveyed WTC during group work. Ghonsooly, Khajavy and Asadpour (2012) and Peng and Woodrow (2010) inspected the components of WTC in general and found some interrelations. Other scholars such as Tseng and Schmitt (2008) directly dealt with motivation and motivation effects on learning. Wolfgramm et al. (2010) dealt with German as a SL conducted with young teenagers in Germany and Switzerland while Csiz´er and Kormos (2009) focused on German and English as a FL in Hungary. Other works concentrated on English as a foreign language in Iran (Papi, 2010), Japan (Hiromori, 2009), Sweden (Henry & Cliffordson, 2009), and the study on vocabulary motivation in Taiwan (Tseng & Schmitt, 2008). To say more, some SEM studies dealt with L2-learning aptitude applying components of the Modern Language Aptitude Test (MLAT). For example, Cochran et al. (2010) examined the way attributions, native-language reading levels, aptitude, and attitudes influenced foreign language learning. Miglietta and Tartaglia (2009) directly explored acculturation elucidating the way parameters like linguistic competence, length of residence, and exposure to media, mediated by practical intra-country application of the target language, e.g., Italian, influenced Italians' feeling of emotional belonging as ESL speakers. Also, Gardner, Tremblay, and Masgoret (1997) centred on the influence of individual differences (IDs) on learning language. Csiz´er and Kormos (2008) focused on the relation between universal posture or general dispositions toward learning language and the understood position of the target language. This work was concerned with the way international posture impacts motivated learning behavior. Rivers (2010) speculated on the manners in which universal posture impacts attitudes toward SL learning. Csiz´er and Kontra (2012) centered on the way various dispositions toward English impinge on people's beliefs concerning English and English-learning purposes.

In testing domain, Mellenberg and Van Der Linden (1982) investigated the choosing items for criterion-referenced tests centering on optimal item selection methods for criterion-referenced tests. Another article (Shrock & Coscarelli, 2010) concentrated on CRT assessment and worked on the major phases in writing a CRT assessment–domain specification, item or task development, validation of content, analysis and selection of items, standard time and place, reliability, and establishing concurrent validity, and described reporting of scores and also the measurement issues pertinent to each. Other scholars delineated numerous issues which emerge when CRT outcomes are applied to examine the impacts of a certain pedagogic intervention concentrating on (1) other substitute methods of combining each student and group data on goals, (2) the sensitivity of the tool to program results, and (3) the comparisons of CRT data and standardized (NRT) achievement test data (Barta,, Ahn, & Gastright, 1976). Furthermore, another study investigated the main applications, problems, and findings of criterion-referenced tests and distinguished three main challenges of CRT assessment: 1. the issue of CRT scoring and score interpretation, 2. the issue of CRT item and test analysis, and 3. the issue of mastery testing (Van der Linden, 1982). Finally, Reed (1992) focused on the association between criterion-based levels of speaking proficiency and NRT scores of overall competency in English as a SL.

As mentioned at the beginning of this section, no study has explored the homogenization of academic English proficiency using structural equation modeling of the correlation between NRTs and CRTs. So, any attempt to integrate and present a systematic analysis of pre-existing knowledge base in this regard will be futile. Rather, this paper will serve as the first base for future endeavors on the topic.

## 3. Method

### 3.1 Participants

The subjects of this research were 210 allied Health EGP students at Shahid Sadoughi University of Medical Sciences in Yazd, central Iran, including 160 junior students of Health and also 50 paramedical students. They were aged 18-20 years, 23 (11%) were male and 187 (89%) were female. All the participants were the junior students of health and paramedicine and were native speakers of Persian. Their English proficiency level varied from –intermediate to +intermediate. They could leave the study at any stage. No subject attrition rate was predicted because all the students were expected to pass the EGP course as a prerequisite for the EAP courses.

### 3.2. Materials and Instruments

#### *3.2.1. Instrument 1*

Cambridge Placement Test (the NRT in this study) including 20 items on reading comprehension, 20 items on grammar, and 20 items on vocabulary was used to place the students at different levels. The “Cambridge Placement Test” is a measure of general English, testing the Reading skill or Use of English and also the Listening skill. It can be used to locate learners at all levels of the “Common European Framework of Reference for Languages (CEFR)” from Pre-A1 to C2. The listening section of the test was omitted since the CRTs had no section on this skill.

#### *3.2.2. Instrument 2*

Also, researcher-made midterm and final exams (CRTs) were used to observe the degree of interrelations among the subcomponents of the

NRT and CRTs. The CRTs items were matched for difficulty level with the NRT using FV (facility value) in a pilot test on 20 subjects. Similar to the NRT, the CRTs included 20 items on reading comprehension, 20 items on grammar, and 20 items on vocabulary (midterm, r=0.76; final exam, r=0.86). The face validity of the test was approved by 10 TEFL faculty members. The content of the test was based on a table of specifications.

### 3.3. Procedure

This study applied SEM to examine the homogenization of academic English proficiency and preferred SEM over other statistics like repeated measures ANOVA because there are some presumed correlations between our independent variables that render other statistics irrelevant. As Kline (2012) asserts, the SEM technique requires large samples, usually $N >$ 200 and the sample size needed somehow depends on how complex the model is, the estimation modality applied, and the distributional features of observed variables. Hence, the research was conducted on 210 allied health students during the 2nd semester, 2017-2018. In this correlational study, convenient sampling method was used to assign 210 subjects into six proficiency groups each including 35 students on the basis of scores on Cambridge Placement Test: two +intermediate, two intermediate, and two –intermediate groups with no control group with the same intervention for all as they were taught with a trained instructor using the EAP-based text: "English for the Students of Medicine I" (Didari & Ziahosseini, 2016) with the same protocol for all groups. A researcher-made 60-item midterm exam (CRT), focusing on the grammar, vocabulary, and reading comprehension, was conducted after the germane evaluation such as content validity based on expert opinions. The exam was identical for all groups. Classes continued up to the end of the semester, and then a researcher-made 60-item final exam (CRT), with the same format as the midterm was conducted. To control the cheating of the students, both multiple choice mid-term and final exams were designed in two parallel forms.

### 3.4. Data Analysis

The related indices for each test such as β-index were calculated. After the final exam, all the data obtained from the Cambridge Placement Test (NRT), the midterm exam (CRT), and the final exam (CRT) were imported into SPSS19 and analyzed to test the first hypothesis using descriptive and inferential statistics such as t-tests, and Pearson Product Moment Correlation. The PLS software was used for $Q^2$ analysis to test the second hypothesis. Then, the correlation matrix was imported to Amos to conduct the assumed SEM used to evaluate the contribution of NRT scores to CRT scores for the six groups in the third hypothesis (Arbuckle & Wothke, 1999). All structural equation modeling analyses were founded on the covariance matrix of the measures. Complementary indices of full latent variable theory were estimated with Amos. Six scores were outliers and were omitted from data analysis. So, the analyses were conducted on the scores of 210 students to avoid confounding of results.

## 4. Results and Discussion

### 4.1. Results

The descriptive statistics of the NRT and CRTs including total scores as well as the scores on the components of three assessments (placement test, midterm exam, and final exam) related to vocabulary, grammar, and reading are displayed in Table 1.

**Table 1**

*Descriptive Statistics of the Total NRTs and CRTs and Their Components (n = 210)*

| Variables | Range | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Total (Placement) | 29 | 10 | 39 | 22.90 | 6.072 |
| Total (Mid-term) | 38 | 18 | 56 | 34.73 | 6.997 |
| Total (Final) | 35 | 19 | 54 | 35.29 | 6.750 |
| Vocabulary (Placement) | 10 | 3 | 13 | 7.06 | 2.268 |
| Grammar (Placement) | 11 | 3 | 14 | 7.03 | 2.093 |
| Reading (Placement) | 18 | 1 | 19 | 8.81 | 3.364 |
| Vocabulary (Mid-term) | 15 | 4 | 19 | 11.15 | 3.116 |
| Grammar (Mid-term) | 21 | 4 | 25 | 11.19 | 3.330 |
| Reading (Mid-term) | 18 | 2 | 20 | 12.46 | 4.255 |
| Vocabulary (Final) | 17 | 2 | 19 | 11.17 | 3.116 |
| Grammar (Final) | 17 | 3 | 20 | 11.09 | 3.178 |
| Reading (Final) | 17 | 4 | 21 | 13.05 | 4.088 |

#### *4.1.2. Testing the First Hypothesis*

The correlational analyses (Table 2) were conducted to analyze the first hypothesis (There exists a statistically significant correlation between NRTs and CRTs as latent variables and their components as observed variables).

The relationships between the total scores of placement tests, midterm, and final exams showed a positive significant relationship between the total score of placement test (NRT) and total score of midterm exam (CRT) ($p$ = 0.0001, $r$ = 0.677), a positive significant relationship between total score of placement test and total score of final exam (CRT) ($p$ = 0.0001, $r$ = 0.721), and also a positive strong significant relationship between total score of midterm exam and total score of final exam ($p$ = 0.0001, $r$ = 0.916).

**Table 2**

*Correlations Matrix of the Components of the NRT and CRTs (n = 210)*

| | Total Place | Total Mid | Total Fin | Voc Place | Gram Place | Read Place | Voc Mid | Gram Mid | Read Mid | Voc Fin | Gram Fin | Read Fin |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TotalPlace | | | | | | | | | | | | |
| TotalMid | .677$^{**}$ | 1 | | | | | | | | | | |
| TotalFin | .721$^{**}$ | .916$^{**}$ | 1 | | | | | | | | | |
| VocPlace | .723$^{**}$ | .475$^{**}$ | .475$^{**}$ | 1 | | | | | | | | |
| GramPlace | .797$^{**}$ | .523$^{**}$ | .555$^{**}$ | .551$^{**}$ | 1 | | | | | | | |
| ReadPlace | .820$^{**}$ | .574$^{**}$ | .635$^{**}$ | .290$^{**}$ | .451$^{**}$ | 1 | | | | | | |
| VocMid | .306$^{**}$ | .577$^{**}$ | .441$^{**}$ | .534$^{**}$ | .126 | .108 | 1 | | | | | |
| GramMid | .351$^{**}$ | .651$^{**}$ | .555$^{**}$ | .185$^{**}$ | .500$^{**}$ | .209$^{**}$ | .246$^{**}$ | 1 | | | | |
| ReadMid | .594$^{**}$ | .706$^{**}$ | .744$^{**}$ | .233$^{**}$ | .383$^{**}$ | .676$^{**}$ | .024 | .168$^{*}$ | 1 | | | |
| VocFin | .334$^{**}$ | .577$^{**}$ | .574$^{**}$ | .482$^{**}$ | .137$^{*}$ | .188$^{**}$ | .649$^{**}$ | .270$^{**}$ | .256$^{**}$ | 1 | | |
| GramFin | .519$^{**}$ | .647$^{**}$ | .710$^{**}$ | .287$^{**}$ | .550$^{**}$ | .402$^{**}$ | .239$^{**}$ | .679$^{**}$ | .367$^{**}$ | .298$^{**}$ | 1 | |
| ReadFin | .532$^{**}$ | .573$^{**}$ | .665$^{**}$ | .195$^{**}$ | .379$^{**}$ | .593$^{**}$ | .051 | .185$^{*}$ | .748$^{**}$ | -.039 | .170$^{*}$ | |

**. Significant correlation at the 0.01 level (2-tailed).

*. Significant correlation at the 0.05 level (2-tailed).

Table 2 shows moderate to strong correlations between nine observed variables of the three latent variable measurement tools, i.e., NRT and CRTs. The strongest correlation was between TotalMid and TotalFin (*p* = 0.0001, *r* = 0.916). Most correlation coefficients suggest a positive correlation between NRT and CRTs confirming the first hypothesis; yet there was no statistically significant correlation between ReadFin and VocMid (*p* = 0.464, *r* = 0.051). Also, there was a negative insignificant correlation between VocFin and ReadFin (*r* = -0.039, *p* = 0.574). This is consistent with our hypothesis because we hypothesized that the components of CRTs are correlated with and predictable from their corresponding components of the NRT.

These correlations indicate that NRT scores can be used to homogenize academic English classes to avoid heterogeneous classes consisting of students of the same major with different language ability levels thereby overcoming the individualistic problems.

### *4.1.3. Testing the Second Hypothesis*

The $Q^2$ (Stone-Geisser Criterion) analysis was run to analyze the second hypothesis of the study (University students' CRTs scores can be significantly predicted based on their NRT scores). If the $Q^2$ values of a dependent construct are estimated to be 0.02, 0.15, and 0.35, this indicates, respectively, the weak, moderate, and strong predictive value of the construct or the independent constructs related to it. In Table 3, the column on the right (1-SSE/SSO) displays the $Q^2$ value of each construct (the mathematical equation is $Q^2$=1-SSE/SSO).

Table 3 shows that all of the values are greater than 0.02 suggesting that the predictive power of none of the constructs is weak. Also, the predictive power of VocFin and GramMid is moderate, while GramFin, ReadFin, TotalFin, VocMid, ReadMid, and TotalMid have a strong predictive value.

**Table 3**

*$Q^2$ (Stone-Geisser Criterion) Analysis*

| Case | SSO | SSE | 1-SSE/SSO |
|---|---|---|---|
| VocFin | 44.311294 | 35.514776 | .198516 |
| GramFin | 32.878320 | 20.115527 | .388183 |
| ReadFin | 41.298907 | 25.169007 | .390565 |
| TotalFin | 36.119440 | 17.615900 | .512288 |
| VocMid | 38.067803 | 21.778379 | .427906 |
| GramMid | 26.912779 | 20.870088 | .224529 |
| ReadMid | 33.865407 | 18.932486 | .440949 |
| TotalMid | 38.304299 | 21.219208 | .446036 |

Note. SSO= sum of squares observed, SSE=sum of squares due to error.

So, we may predict the CRT scores on the basis of the NRT scores confirming our second hypothesis. This means that both CRTs and their components can be predicted on the basis of students’ scores on the placement test. Hence, this prediction will enable academic educators to place university students in linguistically homogenous classes regardless of students’ academic major to avoid problems related to individual differences in proficiency levels.

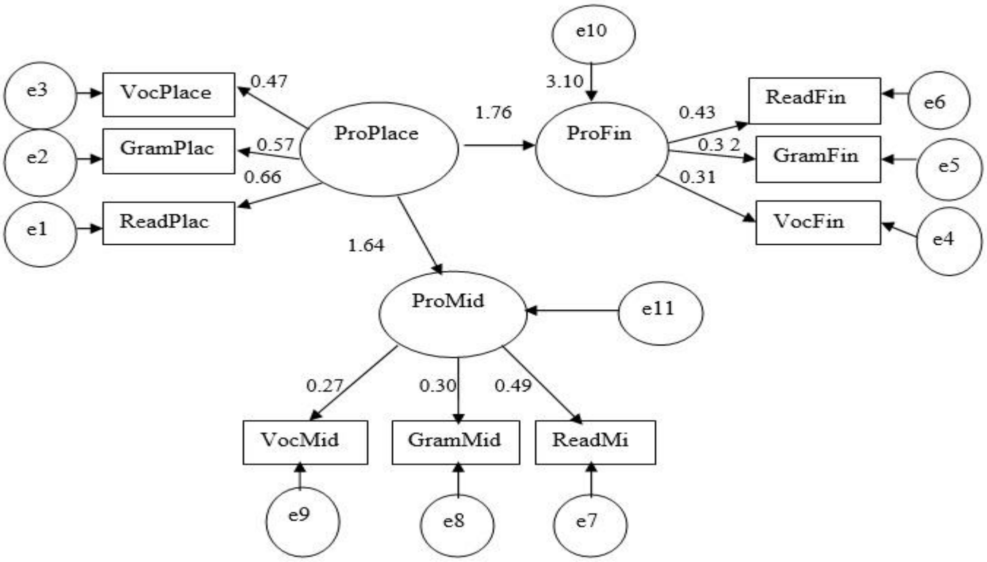
### *4.1.4. Testing the Third Hypothesis*

Following Lei and Wu (2007), our SEM analysis went through the following steps: 1. Specification of Model, 2. Collection of Data, 3. Estimation of Model, 4. Evaluation of Model, and possibly, 5. Modification of Model.

First, the model was specified by determining the latent and observed variables. There were three latent variables including language proficiency at the placement test (LangProfPlace), language proficiency at the midterm exam (LangProfMid), and language proficiency at the final exam (LangProfFin). The scores on the NRT and CRTs were used to indirectly measure these latent variables and were considered as observed variables since language proficiency as a latent variable cannot be directly observed; rather, the latent variables are manifested by the observed scores of the tests on these latent variables. In the Amos model, the latent variables are represented by rectangles and the observed variables are displayed by ovals. The paths between factors and components indicate the correlation between them. These correlations are interpreted as paths, standard estimations, or regression indices. Measurement errors are confined by smaller ovals and indicate the portion of variance accounted for by each observed variable. Then, the data were gleaned as explained in "Procedure" section. Next, having imported the data covariance matrices into Amos, the unstandardized SEM model was estimated using the tests data. Finally, this presumption was confirmed as there were significant paths between the placement test as the NRT and midterm and final exams as the CRTs. As shown in Figure 1, the path coefficient between ProfPlace and ProfMid is significant indicating that language proficiency at the midterm exam is both correlated with and predictable from language proficiency at the placement test. The same is true with the path between ProfPlace and ProfFin. The three paths between each latent variable and its three observed variables are also significant indicating that each observed variable is correlated with its related latent variable. For example, VocPLace, GramPlace, and ReadPlace are both correlated with and predictable from ProfPlace. The same is true with other two latent variables. If a path cannot achieve statistical significance, this does not imply that the predictor factor has no relation to the outcome factor. It may indeed have a significant relationship with the outcome factor. It may also be significantly correlated with other predictor factors of the model, hence diminishing its unique contribution to the forestalling of the result.

**Figure 1**

*Standardized Model*



### *4.1.5. Full Latent Variable Model Fit Assessment*

$R^2$ is a criterion applied to link the measurement model and the structural model of the structural equation modeling. It shows the impact of an independent variable on a dependent one. An important point is that the value of $R^2$ is calculated only for the dependent constructs of the model while the value of this criterion is considered as zero for the independent constructs. The GOF coefficient is used to investigate the total fit of the full latent variable model. According to *Ringle, Wende, and Becker* (2015), GOF values of 0.01, 0.25, and 0.36 conventionally indicate weak, moderate, and strong model fit, respectively. The GOF obtained in this study was 0.381. Regarding the $R^2$ values of the variables displayed in Table 4, the mean of these values equals 0.267 and the mean of the communality values is 1, so the GOF obtained equals 0.517 which indicates the strong fit of the fitted model.

**Table 4**

*AVE & $R^2$*

| Variables | Average | $R^2$ |
|---|---|---|
| VocPlace | .1880453 | |
| VocMid | 1.0000000 | .335424 |
| VocFin | 1.0000000 | .274325 |
| ReadPlace | .1885326 | |
| ReadFin | 1.0000000 | .370112 |
| ReadMid | 1.0000000 | .466037 |
| GramMid | 1.0000000 | .281889 |
| GramFin | 1.0000000 | .332950 |
| GramPlace | .1086214 | |
| TotalPlace | .396900 | |
| TotalMid | 1.0000000 | .459966 |
| TotalFin | 1.0000000 | .526232 |

A common criterion for assessing the fit of default model is Cronbach's-α which is a classical standard for measuring reliability and a suitable measure of internal consistency. Also, a Cronbach's α greater than 0.7 suggests an acceptable reliability (Cronbach, 1951). The value of Cronbach's α was 0.905 indicating the high consistency between each construct and its related index (*Ringle, Wende, & Becker*, 2015).

### *4.1.6. The NFI, CFI, GFI, AGFI, RMR, and RMSEA Indices*

Among the absolute indices, $X^2$ deals with the absolute values of the remainders. The $X^2$ test examines the hypothesis that the desired model is consistent with the covariance pattern among the observed variables. The quantity of $X^2$ is highly dependent on sample volume as a large sample volume increases the $X^2$ value significantly to the degree that it could not be attributed to inaccuracy of the model. Ideally, the $X^2$ should have significance level greater than 0.05. However, some sources have claimed that for the acceptance of the model, the $X^2/df$ should be less than 3 (Kline, 2011). As Table 5 shows, statisticians have not set any criteria with regard to the standard value for $X^2$ and the observed value for our model is 18.19 indicating the good fit of the model.

A large sample volume increases the value of $X^2$ more than it could be attributed to model unfit (Kalantari, 2009; Hooman, 2012). Also, the more correlations there are in the model, the weaker the fit of the model (Kenny, 1979). Ideally, the $X^2$ should have significance level greater than 0.05. However, some sources have claimed that for the acceptance of the model, the $X^2/df$ should be less than 3 (Kline, 2011). These indices do not indicate fit of the model by themselves; rather, they should be interpreted collectively. If the $X^2$ is not statistically significant and the value of $X^2/df$ is less than 2, this indicates the appropriate fit of the model; yet, this index is usually significant in large sample volumes and, hence, is not considered as an appropriate index for assessing model fit. This value is greater than 2 (Table 5). Hence the third hypothesis is confirmed.

The Normalized Fit Index, also called Bentler-Bonett index (Bentler & Bonett, 1980), is acceptable for values greater than 0.9 indicaing good fit of the model. The Comparative Fit Index examines the rate of model improvement through comparing a so-called independence model in effect that there is no correlation between the variables with the default model. The CFI index is similar to NFI index in statistical significance except that it compensates for sample volume. Its value should conventionally equal at least 0.9.

**Table 5**

*Standard Index Values of Model Fit*

| Index | Standard Index Value | Index Value in the default Model | Result |
|---|---|---|---|
| $X^2$ | ---- | 18.17 | Good model fit |
| $X^2/df$ | 1-3 | 2.64 | Good model fit |
| NFI | Greater than .9 | .979 | Good model fit |
| CFI | Greater than .9 | .986 | Good model fit |
| GFI | Greater than .9 (0-1) | .986 | Good model fit |
| AGFI | Greater than .9 (0-1) | .90 | Good model fit |
| RMR | Near zero | .036 | Good model fit |
| RMSEA | Less than .1 | .093 | Good model fit |

The GFI indices assess the relative values of variances and covariances concurrently and jointly through the model. The domain of changes in GFI varies from 0 and 1. The GFI value should be greater than

0.9. Moreover, another fit index is AGFI which is the Adjusted Goodness-of-Fit Index for degree of freedom. The AGFI value should also be ≥ 0.9 for the model to be accepted. Our NFI, CFI, GFI, and AGFI coefficients are greater than 0.9 indicating good fit of the model.

In RMR as another absolute index, the value equals the mean root of squared remainders, i.e., the difference between the observed matrix elements and the estimation matrix elements with the assumption of accuracy of the default model. The closer the RMR of the desired model to zero is, the greater the fit of the model. As displayed in Table 5, the RMR value obtained from our model is 0.036 indicating good fit of the model.

Root Mean Squares of Error Approximations (RMSEA) is one of the absolute fit scales. The RMSEA value should be less than 0.1 for the model to have good fit. This method uses the root mean squares of the difference between the predicted matrix and the observed matrix. Our RMSEA value is 0.093 indicating good fit of the model. Overall, the significant paths explain that CRT scores are correlated with and predictable from NRT scores and share a significant covariance loading.

## 4.2. Discussion

As mentioned earlier, no study has so far investigated the homogenization of academic English proficiency using structural equation modeling of the relationship between NRTs and CRTs. However, in a correlational study by Barzegar et al. (2020), conducted on 71 students of three health majors, the students were assigned into three language ability groups using placement test percentiles. They found a significant disparity among the three groups on the placement test (P=0.015), no significant difference among the three different majors with respect to Criterion-referenced Test (CRT) scores (P=0.05), no significant difference among the three Norm-referenced Test (NRT) forms (Forms A, B, & C) (P=0.05), and a significant difference among the two CRT forms (Forms A, B) (P=0.05). They concluded that the students of health ought to be placed in Basic English and EGP courses, not using their academic majors, rather on the basis of their English proficiency levels for promising EAP teaching. This is consistent with our findings.

In the present study, the first one to use SEM for investigating the effect of homogenization on language proficiency, the default model was fit in the evaluation phase for both paths: one between ProfPlace and ProfMid and the other between ProfPlace and Prof Fin. A significant path originating from a given predictor factor to an outcome factor indicates that the predictor factor possesses unique variance in justifying the outcome factor over and beyond its common covariance with other predictor factors.

The scores on the components of one NRT and two CRTs including vocabulary, grammar, and reading comprehension as observed variables were applied to evaluate the latent variable of language proficiency. The SEM model rendered the paths between the latent variables and their components as significant. As this was the first case of structural equation modeling of the relationship between NRTs and CRTs, there were no studies available to compare and contrast our results with.

Additionally, the relationship between the NRTs and CRTs is an acceptable reality, but presenting a SEM justification would be the innovative aspect of the present study. If the students' future achievements can be predicted on the basis of a standard NRT, educational progress of the students would be guaranteed. Ultimately, a better guidance can be provided for the students.

Our findings suggested that NRTs are correlated to CRTs. Also, $Q^2$ analysis indicated that CRTs are predictable on the basis of NRTs. Moreover, the fitted SEM revealed some significant paths between NRT and CRTs components all suggesting that we can group our students in Basic and EGP courses by homogenizing their language ability using NRT scores of placement tests and overcome the heterogeneous classes in which the university students are placed in different classes using their academic majors, e.g., health, medicine, law, literature, dentistry, etc. A structural model of the correlation between NRT and CRTs can help the language educators to overcome individual differences among the learners. Ehrman et al. (2003) elaborated on the issue of individual differences in language learning. In their viewpoint, this complex topic has meant little conclusive knowledge and thus demands more investigation. Finally, the homogenization of language learners in language centers or institutes is a common practice round the globe. Nonetheless, homogenization of university students' language proficiency in their Basic English, EGP, and EAP courses is an entirely neglected issue. We highlighted this point in this study hoping that it will reduce the challenges that university students face when they are forced to learn English in linguistically heterogeneous classes. This is done through giving placement tests to group them appropriately.

## 5. Conclusion and Implications

Grouping language learners in language institutes on the basis of placement tests is a widespread practice round the globe; yet, placing the university students with different academic majors in linguistically homogenous groups in academic English courses on the basis of NRT sores and predicting their CRT scores on the basis of NRT scores is the innovative aspect of this study. It helps overcome proficiency heterogeneity of academic language classes leading to resolution of problems imposed by individual heterogeneous proficiency levels in classes grouped on the basis of students' academic majors. In this study, we found a positive relationship between scores of NRTs and CRTs as independent observed variables. Also, $Q^2$ index showed that the CRT scores are predictable on the basis of NRT scores. Finally, a structural equation modeling of the association between the NRTs and CRTs indicated significant paths between the components of the NRT as observed variables of a latent variable and the components of CRTs as observed variables of two latent variables. Hence, we should group our university students on the basis of their language ability level and not on the basis of their majors to overcome the heterogeneity of learners' language ability levels and minimize the detrimental effects of individual differences on learning English as a FL/SL. Future studies can focus on structural equation modeling of relationships between general academic English proficiency and the four language skills and components, i.e., vocabulary, pronunciation, syntax, etc.

Future research should focus on larger samples to obtain better results as SEM is sample sensitive, use students of other majors, other proficiency levels, and other variables to design a SEM model to change the language pedagogy in the Iranian setting.

## References

Aidinlou, N. A., & Ghobadi, S. (2012). Examination of relationships between factors affecting on oral participation of ELT students and language development: A structural equation modeling approach. *International Journal of English Linguistics, 2,* 131–141.

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Oxford Journal of Applied Linguistics, 33*(1), 42-65.

Arbuckle, J.L. (2003) Amos 5.0. Smallwaters Corps, Chicago.

Arbuckle, J. L., & Wothke. W. (1999). *Amos 4.0 user's guide.* Chicago: Small Waters.

Aryadoust, V. (2010). Investigating writing sub-skills in testing English as a foreign language: A structural equation modeling study. *TESL-EJ: Teaching English as a Second or Foreign Language, 13,* 1–20.

Barta, M. B., Ahn, U. R., & Gastright, J. F. (1976). Some problems in interpreting criterion referenced results in a program evaluation. *Studies in Educational Evaluation, 2,* 193-202.

Barzegar, K., Nemat Tabrizi, A. R., Jafarigohar, M., & Vahdany, F. (2020). The effect of linguistic homogenization on English proficiency of students, case study: Junior students of Shahid Sadoughi university of medical sciences. *Journal of Community Health Research,* 9(1), 46-53.

Barzegar, K. & Askari, J. (2015). Elucidating idioms through idioms: A metalinguistic contemplation of some issues on "Befogging Idioms". *International Journal of English and Literature,6*(7),109-113.

Bentler, P. M. 1990. Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bond, L. A. (1996). Norm-and criterion-referenced testing. *Practical Assessment, Research & Evaluation*, *5*(2). Retrieved from http://edresearch.org/pare/getvn.asp?v=5&n=2

Cochran, J. L., McCallum, R. S., & Mee Bell, S. (2010). Three A's: How do attributions, attitudes, and aptitude contribute to foreign language learning? *Foreign Language Annals, 43,* 566–582.

Didari, R., & Ziahosseini, M. (2017). *English for the students of medicine I.* SAMT publications, Tehran, Iran.

Ehrman, M. E., Leaver, B. L., & Oxford, R. L. (2003). A brief overview of individual differences in second language learning. *System, 31*(3), 313–330.

Fushino, K. (2010). Causal relationships between communication confidence, beliefs about group work, and willingness to communicate in foreign language group work. *TESOL Quarterly, 44*, 700–724.

Galaczi, E., D. (2014). Interactional competence across proficiency levels**:** How do learners manage interaction in paired speaking tests? *Oxford Journal of Applied Linguistics: Applied Linguistics: 35*(5), 553-574.

Gallagher, H. C. (2013). Willingness to communicate and cross-cultural adaptation: L2 communication and acculturative stress as transaction. *Applied Linguistics, 34*, 53–73.

Gardner, R. C., Tremblay, P. F., & Masgoret, A.-M. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal, 81*, 344–362.

Ghonsooly, B., Khajavy, G. H., & Asadpour, S. F. (2012). Willingness to communicate in English among Iranian non-English major university students. *Journal of Language and Social Psychology*, *31*, 197–211.

Henry, A., & Cliffordson, C. (2013). Motivation, gender, and possible selves. *Language Learning*, *63*, 271–295.

Hiromori, T. (2009). A process model of L2 learners' motivation: From the perspectives of general tendency and individual differences. *System, 37,* 313– 321.

Hooman, H. A. (2012). *Structural equation modeling. Tehran*: SAMT Publications.

Isemonger, I. M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis 2005 and some recommendations for future research in this area. *Studies in Second Language Acquisition, 29,* 101–118.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency**:** How distinct? *Oxford Journal of Applied Linguistics, 29*(1), 24-49.

Jamshidi A., H. & Sadeghi, B. (2013). English language teaching in Iranian universities in brief. *Theory and Practice in Language Studies, 3*(12), 2296-2302.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least square methods*. Mooresville, IN: Scientific Software, Inc.

Kalantari, B. (2009). *Polynomial root-finding and polynomiography*. World Scientific Publishing Co. Pte. Ltd. 5 Toh Tuck Link, Singapore.

Kenny, D. A. (1979). *Correlation and causality* (ch. 3–5). New York: Wiley.

Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, kindergarten to grade 8. *American Educational Research Journal, 48*, 1187–1225.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York, NY: Guilford Press.

Lei, P-W., & Qiong, W. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice, 26,* 33-43.

Lok, B., McNaught, C., & Young, K. (2015). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41(3), 450–465.* doi:10.1080/02602938.2015.1022136.

Miglietta, A., & Tartaglia, S. (2009). The influence of length of stay, linguistic competence, and media exposure in immigrants' adaptation. *Cross-Cultural Research, 43,* 46–61.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Oxford Journal of Applied Linguistics, 24*(4), 492-518.

Papi, M. (2010). The L2 motivational self system, L2 anxiety, and motivated behavior: A structural equation modeling approach. *System, 38,* 467–479.

Peng, J.-E., & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning, 60*, 834– 876.

Phakiti, A. (2008a). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25,* 237–272.

Phakiti, A. (2008b). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly, 5,* 20–42.

Reed, D. J. (1992). The relationship between criterion-based levels of oral proficiency and norm-referenced scores of general proficiency in English as a second language. *System*, *20*(3), 329-345.

*Ringle, C. M., Wende, S., & Becker, J-M. (2015).* Smart PLS 3. *Bönningstedt: SmartPLS GmbH.*

Rivers, D. J. (2010). National identification and intercultural relations in foreign language learning. *Language and Intercultural Communication, 10*, 318–336.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22,* 1–30.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snelling, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing and metacognitive knowledge. *Language Learning, 53,* 165–202.

Shrock, S. A., & W. C. Coscarelli. (2010). Criterion-referenced Measurement. *International Encyclopedia of Education (Third Edition).*

Smemoe, B., W., & Naomi, H. (2013). The effect of language learning aptitude, strategy use and learning context on L2 pronunciation learning. *Oxford Journal of Applied Linguistics,34*(4), 435-456.

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

Song, M.-Y. 2012. Note-taking quality and performance on an L2 academic listening test. *Language Testing*, *29*, 67–89.

Tomiyama, M. (2009). Age and proficiency in L2 attrition: Data from two siblings. *Oxford Journal of Applied Linguistics*, *30*(2), 253-275.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, 58, 357–400.

Tremblay, P. F., & Gardner, R. C. (1995). Expanding the motivation construct in language learning. *The Modern Language Journal*, *79,* 505–518.

Van der Linden, Wim J. (1982). Criterion-referenced measurement: Its main applications, problems, and findings. *Evaluation in Education*, *5*(2), 97- 118.

Wen, Q, & Johnson, R. (1997). L2 learner variables and English achievement: a study of tertiary-level English majors in china. *Oxford Journal of Applied Linguistics*, *18*(1)*, 27-48.*

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics, 34*, 102–122.

Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal, 97*, 109–130.

Wolfgramm, C., Rau, M., Zander-Musi´c, L., Neuhaus, J., & Hannover, B. (2010). On the connection between collective self-worth and the motivation to learn German: A study of students with a migration background in Germany and Switzerland. *Zeitschrift fur Padagogik, 56*, 59–77.

Woodrow, L. J. (2006). A model of adaptive language learning. *The Modern Language Journal*, *90*, 297–319.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing*, *30*, 49–70.

Yang, H.-C. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes*, *31*, 174–187.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*, 558–575.

## Acknowledgements

I thank all the students of allied health for their participation. I also thank Dr. Hossein Fallahzadeh, associate professor in Biostatistics and Epidemiology Department, Amos specialist and Dr. Fatemeh Esfandiary, PhD in statistics and PLS and LISREL specialist for data analysis.

***Bibliographic information of this paper for citing:***

Barzegar, K., Nemat Tabrizi, A. R., Jafari Gohar, M., & Vahdany, F. (2021). English proficiency homogenization of norm-referenced and criterion-referenced tests: A structural equation modeling approach. *Journal of Modern Research in English Language Studies*, *8*(1), 55-75.