Testing Measurement Invariance of an EAP Listening Placement Test across Undergraduate and Graduate Students

Soo Jung Youn
Northern Arizona University
Seongah Im
University of Hawaiʻi at Mānoa

The increasing number of international undergraduates enrolled in English-medium universities creates challenges for an existing EAP (English for Academic Purposes) placement test, especially when the validity of the existing test is not examined with incoming undergraduate examinees. As an attempt to address this issue from a measurement perspective, this study tested measurement invariance in a listening placement test across undergraduate and graduate examinees to investigate whether the test measures the same trait dimension across qualitatively distinct groups of examinees. Using 590 students' listening placement test results, the best fitting baseline model was identified first and then competing models with a series of increasingly restrictive hypotheses were compared to test measurement and structural invariance of the target test across the undergraduate and graduate examinees. Measurement invariance across the undergraduate and graduate examinees was held, indicating invariant factors, equal factor loadings for each item, and error variance. However, structural invariance was not completely established especially for the factor means across two groups, which may suggest different score interpretations and uses depending on examinees' academic status.

Key words: measurement invariance, confirmatory factor analysis, EAP placement test, listening test, M*plus*

Email address for correpsondence: Soo-Jung.Youn@nau.edu

Introduction

In an effort to examine construct validity, researchers advocate investigating the extent to which a language test measures the same construct in the same way across distinct groups of examinees (e.g., Bae & Bachman, 1998; Kunnan, 1998; Shin, 2005). Nonetheless, in reality, it is implicitly assumed that a test works equally for individuals in different subgroups, such as gender or ethnicity. In fact, the establishment of measurement invariance across groups is a logical prerequisite to conducting substantive cross-group comparisons (Vandenberg & Lance, 2000). A lack of measurement invariance across groups reflects the potential incomparability of test results across different groups, indicating the underlying construct measured by the test would not be equivalent and the interpretation of differences or similarities across groups is problematic (Byrne, Shavelson, & Muthén, 1989; Dolan, Oort, Stoel, & Wicherts, 2009; Meredith, 1993).

In the English for Academic Purposes (EAP) context in the U.S., relatively little attention has been paid to how an examinee's academic status (e.g., graduate vs. undergraduate) can affect measurement invariance for EAP placement tests. The student population in an EAP context has changed considerably recently. While the majority of students were graduate students during the last decade in the U.S. (Planty et al., 2009), a noticeable increase in international undergraduate enrollment has been reported (Institute of International Education, 2013; McBride, 2010). Further, the distinct characteristics of undergraduate and graduate students have been reported in terms of varied academic language needs (e.g., Ginther & Grant, 1996; Rosenfeld, Leung, & Oltman, 2001) and different language proficiency levels (e.g., Cho & Bridgeman, 2012). Taken together, establishing measurement invariance for examinees' academic status is essential to ensure that a test functions equally for both undergraduate and graduate students. Thus, the current study investigates measurement invariance of an existing in-house EAP listening placement test to examine the extent to which a factor structure and measurement model of the target test varies across undergraduate and graduate examinees. This way, one can evaluate how an in-house EAP placement test functions across different groups of examinees. To this end, we begin by reviewing the characteristics of undergraduate and graduate students in the field of language testing followed by the discussion of measurement invariance.

Undergraduate and Graduate Students in Language Testing

Inherent differences between international undergraduate and graduate examinees are considered in various assessment contexts. For example, Diagnostic English Language Needs Assessment (DELNA) at the University of Auckland (Elder & von Randow, 2008; Read, 2008) which assesses incoming students' academic readiness has been developed as a result of the considerable numbers of linguistically and culturally diverse undergraduate students. In a large-scale standardized language test context, such as Test of English as a Foreign Language® (TOEFL), group-specific characteristics were examined either at the beginning stage of developing language tests or in post-hoc analyses, such as needs analyses of English use in an academic setting (Ginther & Grant, 1996; Rosenfeld, Leung, & Oltman, 2001) and a difference in L2 proficiency (Cho & Bridgeman, 2012). Rosenfeld, Leung, and Oltman (2001) examined a varying degree of importance of language tasks perceived by graduate students, undergraduate students, and faculty in academic settings to reflect such needs in developing TOEFL items. Although a relatively high correlation (r = .84) was found between the ranks for undergraduate and graduate students' ratings on various academic tasks, slight differences still existed. For example, among various listening tasks in particular, while undergraduate students considered a task of *understanding the instructor's spoken directions regarding assignments and their due dates* the most important, graduate students considered a task of *understanding the main ideas and their supporting information* the most imporant. Further, graduate students' academic language proficiency levels are different from those of undergraduates. Cho and Bridgeman (2012) reported graduate students' higher TOEFL iBT scores than those of undergraduate students, based on a comparison between TOEFL iBT scores of 744 undergraduate and 1,850 graduate students from ten universities in the U.S. The group differences reported in the previous studies may influence ways in which the language ability construct is measured in a language test depending on an examinee's academic status.

Specific to L2 academic listening literature, distinct characteristics between undergraduate and graduate students in terms of different perceptions and needs in their academic listening have been reported. Arguing for the importance of students' perceptions in investigating how stakeholders interpret the genre of L2 academic lectures, Miller (2002) qualitatively analyzed complex factors affecting L2 academic listening and reported that students perceive academic lectures differently. Additionally, Lynch (2011) argues for not considering international students as a single group of listeners, which is supported by Kim's (2006) research findings. Kim reported that East Asian international graduate students' perceptions and needs of L2 academic listening were quite distinct compared to those of undergraudate listeners particularly regarding the difficulty with note-taking skills.

The contexts and research foci of the studies reviewed above vary, but they clearly reinforce the importance of domain or needs analysis reflective of the examinees' characteristics not only in L2 listening tests in particular but also for developing L2 language tests at large. Such attempt is closely related to ensuring the same latent trait being measured across subgroups of population, ultimately warranting valid score interpretations and uses. Considering the potential effect of different examinee characteristics on a test, one can question how an existing in-house EAP placement test functions across undergraduate and graduate examinees. Closer scrutiny of how an examinee's academic status affects a factor structure and psychometric properties of a test will enable us to resolve such a concern.

Measurement Invariance

Measurement invariance, also known as factorial invariance, is often assumed, rather than empirically tested. Yet, the establishment of measurement invariance is a logical prerequiste for examining whether instruments measure the same construct across different groups, hence scores from an instrument can be comparable. Essentially, if measurement invariance holds, an instrument can be applied to individuals at the same construct level across different groups in the same manner. However, models established for each group are not always expected to be identical due to group differences.

Factorial invariance across groups involves two components: *measurement invariance* and *structural invariance* (Byrne, Shavelson, & Muthén, 1989). Measurement invariance can be examined with four different sequential models and structural invariance can be examined with two sequential models thereafter. The two components test distinct psychometric properties. Measurement invariance concerns how items function in measuring a factor across groups and structural invariance addresses how the latent factors are related across groups. While measurement invariance holds, structural invariance may not necessarily hold, indicating real differences in the construct across groups rather than a problem with test instruments. Measurement invariance, as the first component of factorial invariance, concerns the invariance of intercepts, factor loadings, and error variances. A unidimensional factor model expresses the response y_{ijk} to the item *i* as a linear function of the latent factor score θ_{jk} and an error ε_{ijk} of a person *j* in the *k*th group where θ_{jk} ~ N(0,1) and ε_{ijk} ~ N(0,φ_{ik}).

$$y_{ijk} = \mu_{ik} + \lambda_{ik}\theta_{jk} + \varepsilon_{ijk}$$

μ_{ik} is the latent intercept parameter of item *i* in the *k*th group, λ_{ik} is a factor loading and φ_{ik} is the variance of ε_{ijk} for item *i* in the *k*th group where $i=1 \cdots I$, $j=1 \cdots J$, and $k=1 \cdots K$.

Measurement invariance can be investigated with four sequential approaches: (a) *configural*, (b) *metric*, (c) *scalar*, and (d) *strict* invariance. Let us assume we compare factor models for two different groups. *Configural* invariance is to examine whether the same construct(s) is being measured across the two groups. *Metric* invariance concerns the same factor loadings for each item across the two groups, i.e., $\lambda_{j1} = \lambda_{j2}$, in addition to the same construct measured across the two groups. It is also known as weak factorial invariance. *Scalar* invariance, which is strong invariance compared to metric invariance, tests the same item intercepts across groups, i.e., $\mu_{j1} = \mu_{j2}$, in addition to the same construct(s) and same loadings of each item across two groups. Lastly, *strict* invariance tests same error variance of each item across groups, in addition to holding all of configural, metric, and scalar conditions. Accordingly, *configural* invariance is the least restricted and *strict* invariance, testing of equality of all parameters including the error variance, is the most restricted. The second component of factorial invariance, i.e., *structural invariance*, addresses the invariance of factor mean and factor variance-covariance structures across groups thereafter. Factor variances concern how the latent constructs are distributed across two groups, while a factor covariance refers to the degree of relationship between constructs.

Present study

Taken together, given the demographic change in the EAP population, the existing EAP language tests need further examination in order to ensure whether the scores across the undergraduate and graduate examinees can be comparable. The previous research reported the different needs and perceptions of academic listening depending on an examinee's academic status, which can result in different performance on an placement test. Thus, the quality of the existing test toward the incoming undergraduate examinees needs to be examined. Attempts to address these issues can start with examining the structure of the underlying construct being measured in the language test across undergraduate and graduate examinees. The current study aims to test measurement invariance of a listening placement test across undergraduate and graduate examinees using confirmatory factor analyses (CFA). The following research questions guided the study.

1. To what extent is measurement invariance (e.g., a number of factor, factor loadings, item intercepts, error variances) of a listening placement test established across undergraduate and graduate examinees?

2. To what extent is structural invariance (e.g., factor variance, factor mean) of a listening placement test established across undergraduate and graduate examinees?

Method

Participants

The participants in this study were 590 international undergraduate and graduate students at a North American university from five semester test administrations. A recommended minimum sample size for CFA is ten times the number of free model parameters (Raykov & Marcoulides, 2006, 2010). Based on the 30 items used in this study, the number of free parameters ranged from 31 to 91, with a midpoint of 61. The participants' first languages varied, the most common being Chinese, Japanese, and Korean. Of the 590 examinees, 68% were undergraduates ($n = 400$) and 32% were graduate students (n = 190). The undergraduates consist of 260 females (65%) and 140 males (35%); the graduates examinees consist of 102 females (54%) and 88 males (46%). The examinees' TOEFL iBT® scores ranged from 31 and 99 for the undergraduates and from 21 and 105 for the graduates examinees.

Academic Listening Test and Intended Score Use

The target test was an academic listening test as part of a placement test administered for newly admitted international students at an EAP program in a public university in North America, which intends to measure academic listening ability needed to comprehend university-level lectures. The intended use of academic listening test scores was to place incoming international students admitted to a university into a varying level of EAP listening and speaking classes. The placement test development was well documented along with its validity and reliability (Clark, 2007). The listening test consists of five short academic lectures, each includes 5 to 9 questions, totaling 30 multipe-choice items. The items were dichotomously scored (zero for incorrect; one for correct responses). The internal consistency measured by Cronbach's alpha was .751, which is less than the common criterion for good reliability (.80). However, given that only one section (listening test) of the entire placement test was incldued in calculating the reliability, this value is deemed to be acceptable, especially considering a general guideline of acceptable reliability of .70 (Nunnally & Bernstein, 1994). The test was administered under controlled conditions in a large auditorium. All of the instructions and passages for the listening test were delivered via audio recording.

Data Analysis

The latent variable modeling program M*plus* version 6.1 (Muthén & Muthén, 2011) was employed to test measurement invariance of the listening test across the undergraduate and graduate examinees. Since the data consisted of binary responses (i.e., correct vs. incorrect), a weighted least squares (WLS) estimatior was employed with theta parameterization (see Appendix A for an input file code). For the series of analyses to test measurement invariance, several goodness of fit statistics were used to specify how well the hypothesized models fit the data. The goodness of fit of the models was evaluated by using (a) χ^2 test, (b) normed χ^2 (χ^2/df), (c) Comparative Fit Index (CFI), (d) Tucker-Lewis

Fit Index (TLI), and (e) Root Mean Square Error of Approximation (RMSEA). A model typically provides a good fit with the data when the *p*-value associated with a chi-square test is non-significant and χ^2/df is less than three (Wheaton, 1987). The CFI and TLI values greater than .90 indicate adequate fit and values greater than .95 indicate good fit. RMSEA values less than .06 indicate a satisfactory fit (Hu & Bentler, 1999). A change in χ^2 ($\Delta\chi^2$) is a useful metric for comparing nested models (the measurement invariance models described before are sequentially nested); however, χ^2 is sensitive to the sample size. Thus, a change in the CFI (ΔCFI) was also examined since CFI is known as less sensitive to the sample size (Bentler & Bonett, 1980; Brown, 2006; Cheung & Rensvold, 2002; Fan, Thompson, & Wang, 1999; Ullman, 2001).

Results

Descriptive Statistics

Table 1 presents descriptive statistics for the listening test across each subgroup. The means and deviations of the two groups were similar, although the undergraduate students did slightly better than the graduate students. Skewness and kurtosis for each subgroup were within +1 and –1, indicating that the score distributions for each subgroup were normally distributed.

Table 1. Descriptive Statistics

	Mean	SD	Min	Max	Skewness	Kurtosis	*N*
Undergraduate	17.82	4.80	3	30	-0.04	-0.40	400
Graduate	16.12	4.87	6	28	0.26	-0.55	190
All	17.28	4.89	3	30	0.05	-0.52	590

Model Comparison to Test Measurement Invariance

Confirmatory factor analyses with a single latent factor were conducted with M*plus*. As Reise, Widaman, and Pugh (1993) describe, the first step in a multi-group analysis in testing measurement invariance is to freely estimate factor loadings and factor variances in each group, called the baseline model which fits the data in terms of parsimony and meaninfulness. The baseline model assumes that a single latent trait may have accounted for the observed item covariances. The freely estimated model then serves as a benchmark against more restricted models. As a χ^2 test is well known to be sensitive to the sample size, a normed χ^2 (χ^2/df) value was used. As seen in Table 2, the baseline model was plausible with acceptable fit indices (χ^2/df = 1.18, CFI = .959, RMSEA = .017), indicating the 30 listening test items appear to measure the latent construct of academic listening ability.

Next, the baseline model was compared across the different academic status (undergraduate vs. graduate) by testing a series of increasingly restrictive hypotheses. Multigroup CFA models were fit to the data for both undergraduate and graduate groups simultaneously. Four models were tested. Model 1 tested whether there was an equal number of factors, invariant factor loading of each item across groups, and equal item intercepts. Model 2 tested whether there was an equal number of factors, invariant factor loading for all items across groups, and equal item intercepts. Model 3, the most restrictive model, tested whether there was an equal number of factors, invariant factor loading of each item across groups, equal item intercepts, and error variances. In addition to equal error variances, the difference between Model 2 and Model 3 includes whether the factor loadings for all items were invariant (Model 2) or the factor loadings for each item was invariant across groups (Model 3). In order to test structural invariance (factor variance and factor mean), Model 4 further tested whether factor means were zero for two groups, which was added from the most restrictive model (Model 3). All four models tested whether factor variances were one for both groups. The *grouping* option in M*plus* and the DIFFTEST statement were used to compare the nested models.

Table 2 summarizes the results from testing each hypothesis for two groups and Table 3 presents the chi-square difference test results using the DIFFTEST option. In addition, the differences of CFI values (ΔCFI) of the four models were presented in Table 3, as Cheung and Rensvold (2002) recommended using ΔCFI with values higher than .01 as indicative of a significant drop in fit, rather than the $\Delta\chi^2$ which is overly sensitive to the sample size. Except for Model 2, all models tested had acceptable fit indices. Comparing the competing models, Model 1 had better fitting to the data than Model 2 with the significant *p*-value of .000 from the $\Delta\chi^2$, as seen in Table 3. In terms of equal error variances across different academic status, Model 1 and Model 3 were compared. As reported in Table 3, the non-significant *p*-value (.068) of the $\Delta\chi^2$ and the CFI difference of –.007 (less than .01) indicate a more restrictive model (Model 3) was a better fitting model. Although Model 1 had better fitting indices in terms of CFI and TLI than those of Model 3, the more parsimonious Model 3 can be considered as a better one. In order to investigate whether the equal factor means for two groups were plausible, Model 3 and Model 4 were compared. The significant *p*-value (.002) of the $\Delta\chi^2$ and the CFI difference of –.025 (larger than .01) indicate that Model 3 was better fitting. In other words, the restriction in equal factor means across two groups worsened the model fit. In sum, Figure 1 further visualizes Model 3. Only one factor for each group, which is noted as a circle, was found. The factor loadings for item 1, for example (λ_{1U} = .233, λ_{1G}=.218), across the two groups were almost identical. The item intercepts, which are not shown in Figure 1, and error variance, attached to each item (e.g., ε_{1U} *=.946*, ε_{1G} *=.952*), across the groups were almost identical as well. However, in terms of the factor means for each group, the standardized factor mean for the undergraduate group (.000) was higher than the graduate group (–.365). The different factor means indicate that graduate students' listening ability was slightly lower than those of undergraduate students, indicating that structural invariance across the two groups was not completely held.

Table 2. Goodness-of-fit Indices for Competing Models

	Indices					
Model	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA
0. Baseline model	476.116	405	1.18	0.959	0.955	0.017
1. Invariant factors, factor loadings, and item intercepts	940.707	839	1.12	0.929	0.926	0.020
2. Invariant factors, *invariant factor loadings for all items across groups*, and item intercepts	1089.207	868	1.25	0.846	0.845	0.029
3. Invariant factors, factor loadings, item intercepts, and *error variances*	979.323	868	1.13	0.922	0.922	0.021
4. Invariant factors, factor loadings, item intercepts, error variances, and *equal factor mean*	1006.386	869	1.16	0.904	0.904	0.023

Table 3. *χ^2 Test for Difference Testing*

	$\Delta\chi^2$	Δ *df*	*p*	ΔCFI
Model 1 vs. Model 2	98.454	29	0.000	-0.083
Model 1 vs. Model 3	41.062	29	0.068	-0.007
Model 3 vs. Model 4	57.474	30	0.002	-0.025

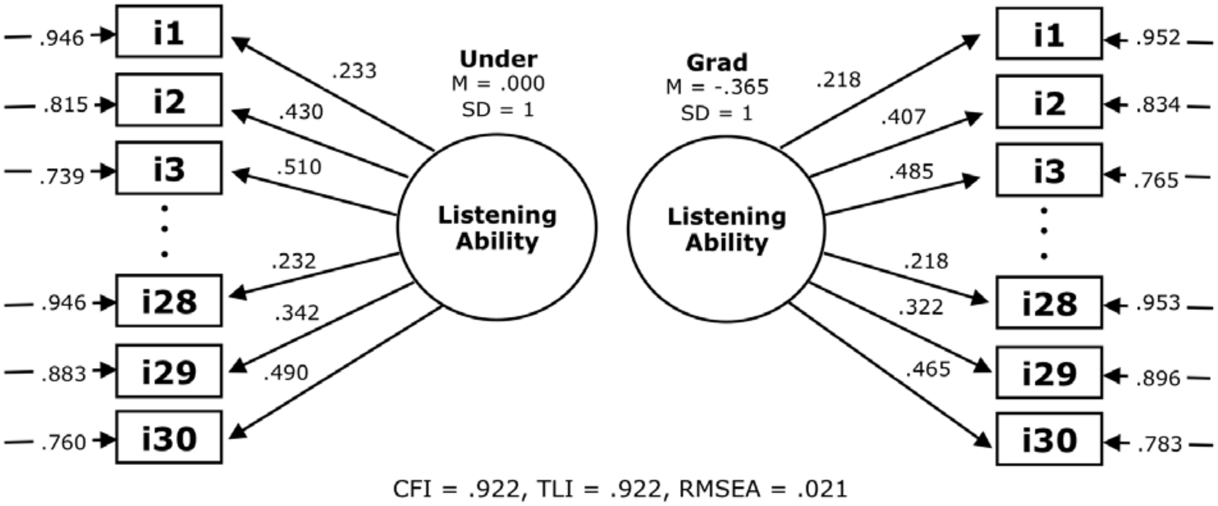

Figure 1. Model 3 Invariant factors, factor loadings, item intercepts, and error variances

Discussion

This study tested measurement invariance of the academic listening placement test, often implicitly assumed as a component of construct validity, across qualitatively distinct two groups of examinees (undergraduate and graduate students) in an EAP context. The plausibility of the single confirmatory factor model (i.e., Baseline model) confirmed that one construct (i.e., academic listening ability) was being measured in the listening placement test. The four models with varying degrees of restrictions were tested and compared based on the basline model. Each model tested increasingly restrictive hypotheses using the χ^2 difference tests. Model 1 tested whether there was an equal number of factors, invariant factor loading of each item across groups, and equal item intercepts, while Model 2 tested whether the factor loadings for all items are invariant along with an equal number of factors and equal item intercepts. Model 3 added another restriction to Model 1, equal error variances. Model 4 tested equal factor means across groups to test structural invariance. All four models tested whether factor variances were one for both groups. Among the four models, Model 3 was found to be the best fitting model, suggesting there was an equal number of factors (i.e., single factor), invariant factor loadings for each item, equal item intercepts, and equal error variances across the undergraduate and graduate examinees. However, Model 4 showed poorer fit than Model 3, which indicates two groups had different factor means. Taken together, in terms of the psychometric structure of the academic listening test, measurement invariance was established. However, structural invariance was not completely held, suggesting the evidence for the differences in two subgroups' academic listening ability.

Our aim for the present study was to examine if the listening test administered at a university for both undergraduate and graduate students would invariantly measure their listening ability across the two groups. The sequential comparisons of measurement invariance modeling indicate that the test holds the strict invariance, the highest level of measurement invariance (i.e., invariant factors, factor loadings, item intercepts, and error variance). We conclude that the listening test items functioned equally for both graduate and undergraduate students, implying that the listening placement test does not either advantage or disadvantage either group of examinee from the perspective of measurement invariance. Yet, the best fitting model, Model 3, also suggested that the standardized factor mean for the graduate examinees was smaller (–.365) than the undergraduate examinees (standardized factor mean of .000), indicating the graduate examinees' lower performances on the listening test. Considering the graduate students' higher L2 proficiency levels reported in previous research (e.g., Cho & Bridgeman, 2012), the undergraduate examinees' out-performance on the academic listening test in this study is rather unexpected. At the same time, this finding further suggests that graduate students' out-performances cannot be taken for granted. Although reasons for the graduate examinees' lower performances cannot be directly inferred from this study, this finding reflects potentially distinct characteristics of undergraduate examinees, supporting Lynch's (2011) argument that international graduate and undergraduate students are distinct listener groups. Further, the different factor means reported for the two subgroups suggest that the findings could have an impact on the test use. It might be necessary to reevaluate the appropriateness of cut scores used to place undergraduates and graduate students into remedial classes. At the same time, while a psychometric structure in terms of measurement invariance was identical across the two subgroups in this study, qualitatively distinct characteristics still deserve further attention.

Conclusion

The current study was motivated by the recent surge of incoming international undergraduate students in English-medium universities and its potential effect on an existing placement test from a measurement perspective. Thus, rather than assuming a test works equally for different subgroups, the primary goal of the present study was to examine the quality of the existing listening placement test focusing on measurement invariance across the undergarduate and graduate examinees with empirical evidence. In general, measurement invariance of the listening test was established across the undergraduate and undergraduate examinees in terms of a number of factors being measured, factor loadings, item intercepts, and error variances. The fact that the academic listening placement test functioned equally for the qualitatively distinct two groups is quite inspiring and important, which adds strong evidence for the quality of the test developed for the university students and reassures stakeholders in administering the test toward incoming undergraduate students. Additionally, the differing factor means (i.e., listening ability) suggest more careful attention toward using scores in placing each group of examinees rather than treating the examinees as one identical group.

The finding that the listening placement test equally functions for the two different groups of examinees demonstrated a subtle psychometric feature of the test. Further attention on improving the moderate degree of reliability of the listening placement test used in this study is necessary as future research. The limitation with regard to the sample size should be noted when it comes to interpreting this finding. Although an equal sample size is not required for measurement invariance across diffferent groups, the sample size for the undergraduate examinees was almost twice as large as the sample size for the graduate examinees in this study, which might have influenced the finding. Finally, as the EAP student population becomes more diverse, other qualitatively distinct groups of examinees, such as L1 and gender, deserve future research in investigating measurement invariance of in-house EAP tests.

References

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*, 380–414.

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

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

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*, 421–442.

Clark, M. K. (2007). *Listening placement test development and analysis from a Rasch perspective* (Doctoral dissertation). Retrieved from Dissertations and Theses database. (UMI No. 3264845).

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling, 16*, 295–314.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool, *Language Assessment Quarterly*, *5*, 173–194.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6,* 56–83.

Ginther, A., & Grant, L. (1996). *A review of the academic needs of native English-speaking college students in the United States* (TOEFL Monograph Series No. 1). Princeton, NJ: Educational Testing Service.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria in fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Institute of International Education (2013). *Open doors 2013: International students in the United States and study abroad by American students are at all-time high*. Retrieved from http://www.iie.org/Who-We-Are/News-and-Events/Press-Center/Press-Releases/2013/2013-11-11-Open-Doors-Data.

Kim, S. (2006). Academic oral communication needs of East Asian international graduate students in non-science and non-engineering fields. *English for Specific Purposes*, *25*, 479–489.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing*, *15*, 295–332.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*, 79–88.

McBride, K. (2010, November 15). UI international undergraduate enrollment surges, study abroad continues to grow. *The University of Iowa News Services*. Retrieved from http://news-releases.uiowa.edu/2010/november/111510open-doors.html.

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

Miller, L. (2002). Towards a model for lecturing in a second language. *Journal of English for Academic Purposes*, *1*, 145–162.

Muthén, B., & Muthén, L. K. (2011). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Planty, M., Hussar, W., Snyder, T., Kena, G., KewalRamani, A., Kemp, J., Bianco, K., & Dinkes, R. (2009). *The condition of education 2009* (NCES 2009-081). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from http://nces.ed.gov/programs/coe/indicator_ins.asp.

Raykov, T., & Marcoulides, G. A (2006). *A first course in structural equation modeling* (2nd ed.). New York: Psychology Press.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York: Routledge.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, *7*, 180–190.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.

Rosenfeld, M., Leung, P., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. 21). Princeton, NJ: Educational Testing Service.

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

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using Multivariate Statistics* (4th ed., pp. 653–771). Needham Heights, MA: Allyn & Bacon.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Method, 3,* 4–70.

Wheaton, B. (1987). Assessment of fit in onveridentified models with latent variables. *Sociological Methods and Research*, *16*, 118–154.

Appendix A: M*plus* input file code

```
DATA: FILE IS c:\data\status 2groups.dat;
        Format is free;
VARIABLE: NAMES ARE i1-i30 academic;
        CATEGORICAL ARE i1-i30;
        Usevariables i1-i30;
        grouping is academic (1=Under 2=Grad)
ANALYSIS: PARAMETERIZATION=THETA;
        DIFFTEST IS deriv.dat;
MODEL:  f by i1* i2-i30;
        f@1;
        i1-i30(1);
OUTPUT: TECH1 standardized;
PLOT: TYPE = PLOT3;
```