Listening. Learning. Leading.®

Research Report
ETS RR–15-25

Investigating the Relationship Between Test-Taker Background Characteristics and Test Performance in a Heterogeneous English-as-a-Second-Language (ESL) Test Population: A Factor Analytic Approach

Venessa F. Manna

Hanwook Yoo

December 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Investigating the Relationship Between Test-Taker Background Characteristics and Test Performance in a Heterogeneous English-as-a-Second-Language (ESL) Test Population: A Factor Analytic Approach

Venessa F. Manna & Hanwook Yoo

Educational Testing Service, Princeton, NJ

This study examined the heterogeneity in the English-as-a-second-language (ESL) test population by modeling the relationship between test-taker background characteristics and test performance as measured by the *TOEFL iBT*® using a confirmatory factor analysis (CFA) with covariate approach. The background characteristics studied included: (a) main reason for taking the TOEFL iBT test; (b) time spent studying English; (c) time spent attending a school, college, or university in which content classes were taught in English; and (d) lived in a country where English is the main language. The results indicated that at most levels of the background characteristics studied, there were statistically significant differences in the means of the four underlying latent factors (reading, listening, speaking, and writing) representing English-language proficiency (ELP). Overall, the effect size differences on the reading, listening, speaking, and writing latent factors among the levels for each of the background variables studied ranged from small to medium. The results of this study provide empirical evidence of the association and possible influence of test-taker background characteristics on the four underlying latent factors representing ELP and, thus, on test performance.

Keywords MIMIC; population heterogeneity; factor structure; *TOEFL iBT*®

doi:10.1002/ets2.12072

Test validation is the process of making an argument for the proposed interpretation and uses of test scores by stating a series of propositions and collecting evidence in support of these propositions (Kane, 2006). In the context of language testing, one of the essential propositions is that the structure of the test is consistent with theoretical views of language ability. As stated by Bachman (1990), "[A] clear and explicit definition of language ability is essential to all language test development and use" (p. 3), and empirical evidence is central in providing support for these theoretical views.

The theoretical view of language ability and empirical evaluation of these views have been and continue to be the theme of much discussion in the field of language testing. As noted by Bachman (2000), Oller's (1976, 1979) *unitary trait hypothesis* of language ability had the greatest impact on the language-testing field in terms of "repercussions and subsequent research" (p. 22). This hypothesis was in contrast to the four-skill approach (reading, listening, speaking, and writing) proposed by Carroll (1965). However, Oller's unitary trait hypothesis was criticized based on methodological flaws, and in 1983, Oller withdrew his claim that language ability consists of a single unitary ability. Since then, researchers have proposed and empirically evaluated several competing models of language ability that include two or more language abilities that are distinct, closely correlated, or hierarchically related to a global ability (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Bachman & Palmer, 1982; Bae & Bachman, 1998; In'nami & Koizumi, 2012; Kunnan, 1995; Sawaki & Sinharay, 2013; Sawaki, Stricker, & Oranje, 2008; Shin, 2005).

Although researchers have concluded that multiple components comprise language ability (Carroll, 1965; Oller, 1983), no consensus exists on the exact nature of the relationship among the components as manifested across and within language tests (Kunnan, 1998; Powers, 2010; Sawaki et al., 2008; Wolf et al., 2008). This finding is not surprising, as the research findings are influenced by the assessment instruments (which tend to assess or emphasize different subskills) and methodological factors. For example, Sawaki et al. (2008) investigated the factor structure of the *TOEFL iBT*® test using item-level confirmatory factor analysis (CFA). In this study, researchers identified a higher-order general factor to

Corresponding author: V. F. Manna, E-mail: VManna@ets.org

represent English-as-a-second-language (ESL) or English-as-a-foreign-language (EFL) ability and four first-order factors as reading, listening, speaking, and writing. In a subsequent study, Stricker and Rock (2008) confirmed the invariance of this higher-order factor structure underlying the TOEFL iBT test for three subgroups: (a) Indo-European (IE) and non-Indo-European (NIE) language families, (b) Kachru's (1984) outer and expanding circle of countries based on the prevalence of English use in educational and business context, and (c) years of classroom instruction in the English language. However, Stricker and Rock noted that these findings do not rule out the potential influence of these variables on test-taker performance on the TOEFL iBT test, as researchers (as well as Sawaki et al.) used data from participants recruited for the field trial of the TOEFL iBT test. Thus, the sample used for the study may not be representative of the current test population in terms of educational and other background characteristics. In a more recent study, Sawaki and Sinharay (2013) used data from operational administrations of the TOEFL iBT test and found that the correlated four-factor structure, corresponding to the four language skills (reading, listening, speaking, and writing), best represented the latent structure of the test. In contrast to Sawaki et al.'s study, Sawaki and Sinharay used an item-parcel approach (items grouped by content category) rather than item-level approach to the CFA. Although the findings from these studies do not show agreement on the exact nature of the latent structure of the TOEFL iBT test, they are consistent with a theoretical view of language proficiency as a set of highly interrelated components. The findings also show the methodological differences that prevent researchers from reaching a consensus empirically on the components that constitute language ability and the relationship among these components.

The influence of test-taker background characteristics on the language ability has also been the focus of extensive research in the field of language testing (Kunnan, 1998). As part of his communicative language ability model, Bachman (1990) argued that test-taker background characteristics comprise one of three primary factors that affect performance on language tests. These background characteristics include cultural background, background knowledge, field dependence, native language, cognitive ability, gender, and age (Bachman, 1990; Kunnan, 1998). Bachman (2000) restated the significance of test-taker characteristics on performance and scores on language tests. Gradman and Hanania (1991) used an extensive list of 44 background characteristics in their study of the language-learning background characteristics that had the greatest effect on ESL students' language proficiency as measured by their *TOEFL*® test scores. Based on the study findings, the researchers concluded that additional reading outside of the classroom played a prominent role in language learning. Additional factors that contributed to increased language proficiency included exposure to teachers who are native English speakers, the use of English as a language of instruction, and participation in intensive language programs.

Each of these four language skills — reading, listening, speaking, and writing — though strongly correlated, are distinct, and thus, test-taker background characteristics may have a differential impact on these skills. Bae and Bachman (1998), for example, investigated the reading and listening performance on a Korean language test that included both academic and general language tasks in an elementary school setting with two groups: heritage Korean speakers and nonnative speakers of Korean. This study indicated that the correlated two-factor model provided the best fit to the data for both groups; however, variations were evident in the reading and listening proficiencies across groups. The heritage learners showed less variability for listening, whereas nonnative speakers of Korean showed less variability on reading. In another study, Wilson (2000) showed that performance on listening comprehension as measured by the *TOEIC*® test varied more with English use/exposure in comparison with educational level. The opposite pattern was observed on the reading measure, where test takers with less than university level education had relatively lower means compared with higher educational categorization.

In a recent study, Gu (2014) examined the latent components of academic English-language ability as measured by the TOEFL iBT test and differences in means on these latent components for two groups of English-language learners: learners with exposure to an English-speaking environment and nonexposure learners. Gu found that the test-taker performance on this four-skill test could be represented by a correlated two-factor model with one factor representing reading, listening, and writing ability and the other factor representing speaking ability. The findings from this study indicated that the latent structure was invariant across the two groups of English-language learners. Moreover, the study showed that exposure to an English-speaking environment had no impact on the latent factor means and researchers concluded that the development of language ability was comparable in the two groups of English learners. However, the sample size was only 370 test takers across the two groups, and because of the latent structure chosen (reading, listening, and writing ability vs. speaking ability), it was not possible to validate the differential impact of test-taker background characteristics on reading and listening skills found in other studies (e.g., Bae & Bachman, 1998; Kunnan, 1995).

Another relevant line of research is the variation in impact of the background characteristics associated with test takers' native language family. As pointed out by Kunnan (1995), the influence of background characteristics, such as prior exposure to the English language, on test performance can be substantial in some cases, depending on native language group (IE vs. NIE). In Kunnan's study, prior exposure to English was represented by several variables, including home country formal instruction, home country informal exposure, English-speaking country instruction, and English-speaking country exposure.

Research into the influence of test-taker background characteristics, such as prior exposure to English language or native language group, on English-language proficiency (ELP) is not entirely new. However, recent studies have been limited by small sample size, the specific groups studied, and the environmental/educational factors explored. In addition, the nature of the target population of interest is shifting. English is becoming increasingly important as a lingua franca; that is, it is commonly used as a medium of instruction and as a means of communication among nonnative speakers of English globally. Moreover, access to English is more widespread because of the ubiquitous availability of technology. As a result, there is a continuing increase of diverse background characteristics in the ESL/EFL population.

It is important to continue the tradition of adding to the existing body of validity evidence underlying components of ELP by conducting studies with test takers who represent the diversity of the current ESL/EFL population. Hence, the motivation for this study, the purpose of which was to extend the current body of research into the heterogeneity of the ESL/EFL test population by examining the relationship between test-taker background characteristics and test performance as measured by the TOEFL iBT test.

At test registration, TOEFL iBT test takers respond to a questionnaire designed to collect information on test takers' background that ranges from time spent studying English to type of institution the test taker is interested in attending. This study was focused on the following subset of background questions hypothesized to influence ESL/EFL test performance:

1. What is your reason for taking the TOEFL iBT test?
2. How much time have you spent studying English?
3. How much time have you attended a school, college, or university in which content classes (such as mathematics, history, or chemistry) were taught in English?
4. Have you lived in a country where English is the main language?

The first question was used to study the impact of test usage and associated stakes on ELP as measured by the TOEFL iBT test. It was hypothesized that higher levels of ELP are expected when test takers use test scores for higher education application or professional achievement. The remaining three questions were designed to elicit information on test takers' prior exposure to English. Overall, it was hypothesized that test takers will exhibit higher levels of ELP as measured by the TOEFL iBT test with greater exposure to English, although there may be differential performance on the language modalities, depending on whether the exposure was in an academic or nonacademic environment.

Also examined in this study was the differential impact of background characteristics on ELP associated with test takers' native language family. The IE and NIE language family groupings used by other researchers (Ginther & Stevens, 1998; Kunnan, 1995; Shin, 2005; Stricker & Rock, 2008; Swinton & Powers, 1980) were adopted. The languages in the IE family, including English, share common lexicon, phonological, and morphological characteristics. Thus, it is hypothesized that the background characteristics measuring exposure to English would have a greater positive impact on test takers in the same language family as English, that is, the IE language family, than the test takers in the NIE language family.

Method

Data

The data for this study were the scored item responses from 22,624 test takers in the four sections of a TOEFL iBT test form administered in the fall of 2013. The sample included test takers from Europe (18%), the Americas (14%), the Middle East (7%), Africa (1%), China (29%), Korea (10%), India (7%), Japan (5%), and other Asian countries (9%). There were slightly more males in the sample (52%) than females (48%). The performance of the test takers in the sample on the four sections (reading, listening, speaking, and writing) and the total across the sections are summarized in Table 1. Also shown in Table 1 is the performance of all test takers who were tested in 2013 (Educational Testing Service, 2013). For each section, scale scores that range from 0 to 30 were reported. The total reported scale score was the sum of the section

Table 1 Comparison of Test-Taker Performances Between Sampled Data and 2013 Population

		Current sample		All 2013 administrations	
Section	Scale score range	*M*	*SD*	*M*	*SD*
Reading	0–30	21.5	6.3	19.7	6.7
Listening	0–30	21.0	6.5	20.1	6.7
Speaking	0–30	20.8	4.5	20.1	4.6
Writing	0–30	21.3	4.6	20.6	5.0
Total	0–120	84.6	19.3	81.0	20.0

scores and ranged from 0 to 120. A comparison of the summary statistics in Table 1 indicates that the sample used in this study is slightly higher performing, on average, than the general population of 2013 examinees.

Structure of the Test

Tasks in the TOEFL iBT test are designed to provide evidence of academic language proficiency in the four language modalities of reading, listening, speaking, and writing. The speaking and writing sections contain tasks that assess integrated skills across two or more modalities. The reading section includes three item sets, each of which contains 14 items associated with an academic reading passage of approximately 700 words. Each reading item is dichotomously scored, with the exception of the last item in each item set, which is polytomously scored on a 0–2 scale. The listening section includes four academic lectures and two conversations. Each lecture is associated with six items, while each conversation has five items to which test takers respond. All items in the listening section are dichotomously scored.

The speaking section contains six tasks. Two of the speaking tasks are independent; that is, no oral or written stimulus materials are associated with the tasks. Four of the speaking tasks are integrated tasks. In two of the integrated tasks, test takers are required to read a passage, listen to a spoken text that pertained to the reading text, and then respond orally to what was read and heard. On the other two integrated tasks, test takers first listen to a short, spoken conversation then respond orally to what they heard. Each task is polytomously scored on a 0–4 scale. The writing section contains two tasks. The first task is an integrated task in which test takers are required to read a text, listen to a lecture that pertained to the topic, and then write a summary on what they read and heard. The second task is an independent task in which test takers are required to write an essay on a common topic based on their experience and knowledge. Both tasks are scored on a 0–5 scale in increments of .5.

Background Variables

Table 2 shows the four background questions used in this study and the categories to which test takers responded. For analysis purposes, the categories were redefined to include a sufficient number of cases in each category. Also shown in Table 2 are the number and percent of test takers in each response category. The percentages shown were based on the number of test takers who responded to the question. The response rate for the question on the reason for taking the TOEFL iBT test was 97%. The response rates were low for the remaining three questions, with rates of 50%, 48%, and 48%, respectively.

To aid in the interpretability of the results, the response categories for the three language exposure questions were dichotomized in the study of the interaction of test takers' native language with prior exposure to English. In terms of both statistical significance and interpretability, the most meaningful levels of response category from each question were chosen to set the dichotomous level of language exposure. Then, test takers (representing a total of 154 languages) were classified into language groups, IE or NIE, according to their language family, using the criteria implemented in other research studies (e.g., Kunnan, 1995; Stricker & Rock, 2008). The main languages included in the IE group were Dutch, French, German, Hindi, Portuguese, and Spanish, whereas the NIE group mainly consisted of Arabic, Chinese, Japanese, and Korean. The revised response categories and the number and percent of test takers for the dichotomized covariates based on language family are shown in Table 3.

Table 2 Background Questions and Percent Response per Category

Question	Actual response categories	Redesigned response categories	Percent (*N*) response
How much time have you spent studying English?	None	Less than 2 years (reference group)	9% (1,052)
	Less than 1 year		
	1 year or more, but less than 2 years		
	2 years or more, but less than 5 years	2 years or more, but less than 5 years	15% (1,678)
	5 years or more, but less than 10 years	5 years or more, but less than 10 years	37% (4,272)
	10 years or more	10 years or more	39% (4,506)
How much time have you attended a school, college, or university in which content classes (such as math, history, or chemistry) were taught in English?	None	None (reference group)	24% (2,602)
	Less than 1 year		
	1 year or more, but less than 2 years	Less than 2 years	22% (2,387)
	2 years or more, but less than 3 years	2 years or more, but less than 5 years	20% (2,205)
	3 years or more, but less than 5 years		
	5 years or more, but less than 10 years	5 years or more, but less than 10 years	16% (1,787)
	10 years or more	10 years or more	18% (1,924)
Have you ever lived in a country where English is the main language?	No	No (reference group)	53% (5,806)
	Yes, for less than 6 months		
	Yes, for 6 months to 1 year	Less than 1 year	25% (2,772)
	Yes, for more than 1 year but less than 2 years	1 year or more, but less than 5 years	16% (1,718)
	Yes, for 2 years or more, but less than 3 years		
	Yes, for 3 years or more, but less than 5 years		
	Yes, for 5 years or more, but less than 10 years	5 years or more	6% (714)
	Yes, for 10 years or more		
What is your reason for taking the TOEFL test?	Attend secondary school (high school)	High school/community college	8% (1,677)
	Attend a 2-year college/community college		
	Attend an undergraduate program	Undergraduate	31% (6,744)
	Attend a (post)graduate nonbusiness program	Graduate/postgraduate (reference group)	49% (10,865)
	Attend a (post)graduate business program		
	Attend an English-language school or program	Licensure or certification	3% (737)
	For licensure or certification	Employment/immigration	2% (416)
	For employment or job		
	For immigration purposes		
	Other	Other	7% (1,575)

Table 3 Background Questions With Language Family Interaction and Percent Response per Category

Information for language family	Redesigned response categories	Percent (*N*) response
Language family and time spent studying English	NIE and less than 5 years (reference group)	14% (1,519)
	NIE and 5 years or more	41% (4,625)
	IE and less than 5 years	10% (1,067)
	IE and 5 years or more	36% (4,007)
Language family and time spent attending content classes taught in English	NIE and less than 5 years (reference group)	35% (3,813)
	NIE and 5 years or more	18% (1,977)
	IE and less than 5 years	31% (3,353)
	IE and 5 years or more	15% (1,613)
Language family and lived in a country where English is the main language	NIE and less than 1 year (reference group)	40% (4,305)
	NIE and 1 year or more	14% (1,520)
	IE and less than 1 year	39% (4,218)
	IE and 1 year or more	8% (821)

Note. NIE = non-Indo-European language family; IE = Indo-European language family.

Analyses

To address the research questions of interest, a CFA with covariate or multiple indicators and multiple causes (MIMIC) model approach (Jöreskog & Goldberger, 1975; B. O. Muthén, 1989) was used to study the relationship between the background characteristics of interest and the latent factors. The advantages of using the MIMIC approach rather than multiple-group CFA (MG-CFA) in this study were its parsimony (with fewer freely estimated parameters) and smaller sample size requirement. The background characteristics in this study had four to six groupings for comparison. Thus, the MG-CFA would be very cumbersome given complexity of the measurement models being tested and the number of parameters to be estimated and held constant across groups. The MIMIC model assumes (a) the same factor loadings and observed residual variances/covariances for all levels of the covariates and (b) same factor variances and covariances for all levels of the covariates. Two steps were involved in modeling these relationships using the MIMIC process. First, using the full sample, a viable CFA measurement model was established. This step was then followed by the addition of the background variable of interest to the CFA model to examine its direct effect on the latent factors of the CFA model. If there is a significant effect of covariate on the latent factors, this implies the population heterogeneity; that is, the factor means are different depending on the levels of the covariate.

In the first step, an item-level CFA was used to evaluate several competing models representing the factor structure of the TOEFL iBT test. These models used in this study were similar to those studied in previous CFA analyses of the TOEFL iBT test (Gu, 2014; Sawaki & Sinharay, 2013; Sawaki et al., 2008; Stricker & Rock, 2008). Specifically, the following six factor structures, as described in Sawaki et al. (2008), with the exception of 4a, were evaluated and are diagramed in Appendix.

1. Bifactor model (Figure A1): This model hypothesized the presence of a general factor (e.g., ELP) that loads directly on the observed variables as well as four skill factors corresponding to reading, listening, speaking, and writing.
2. Correlated four-factor model (Figure A2): This model hypothesized the presence of four psychometrically distinct but correlated factors defined by the items assessing each of the four language skills. This model is nested within the bifactor model.
3. Single-factor model (Figure A3): This model hypothesized that the four language skills are psychometrically indistinguishable from each other and that a single factor underlies the test.
4. Correlated two-factor models (Figures A4 and A5): Two versions of this hypothesized two-correlated factors model were studied:
    a. In the first version, one factor was defined by the reading and listening items and the second factor by the speaking and writing items.
    b. In the second version, the reading, listening, and writing items defined one factor, and the speaking items defined the other factor.
5. Higher-order factor model (Figure A6): This model is similar to the bifactor model in that four factors are defined by the four language skills and a general factor. However, unlike the bifactor model, the general factor is defined by the latent factors.

All analyses were conducted with MPLUS 7.11 (L. K. Muthén & Muthén, 2012). The weighted least-squares with mean and variance adjustment (WLSMV) were used in the estimation of model parameters. The variance of the factor loadings were set to 1 for scale identification. The model evaluation process began with a choice of a baseline model, against which competing models were evaluated. This evaluation involved a comparison among the two least-restrictive models, the bifactor, and correlated four-factor models. Several goodness-of-fit indices were used to evaluate the model-to-data fit.

This evaluation included the root mean squared error of approximation (RMSEA) and its 90% confidence interval (CI), with a value of .05 or below indicating a good fit (Browne & Cudeck, 1993). Two additional fit indices used were the comparative fit index (CFI) and the Tucker-Lewis index (TLI). Hu and Bentler (1999) recommended a cutoff value close to .95 for both of the indices. Although chi-square test of model fit (χ^2) was provided, the sample size used in this study was large; thus, the chi-square difference test was not used for model comparisons. These statistics are sensitive to sample size, and trivial departures from fit will yield a significant chi-square (Fabriger, Wegener, MacCallum, & Strahan, 1999). In addition to comparison of goodness-of-fit indices, model parsimony and interpretability, the reasonableness of individual parameter estimates (i.e., statistical significance, residuals, and modification indices), and correlations among the latent factors (i.e., correlation higher than .90 as extreme) were also considered as criteria in the model evaluation.

After a viable CFA model was chosen to represent the factor structure of the test, the covariates were then added to the CFA model. This structure is diagramed in Figure 1, assuming that a correlated four-factor model represents the latent structure of the test. It should be noted that it is possible to add several covariates at once to the model. However, in this study, each of the covariates was added and evaluated independently. All the covariates used in this study were categorical. Thus, dummy codes were used to represent group membership. For example, for the covariate time named spent studying

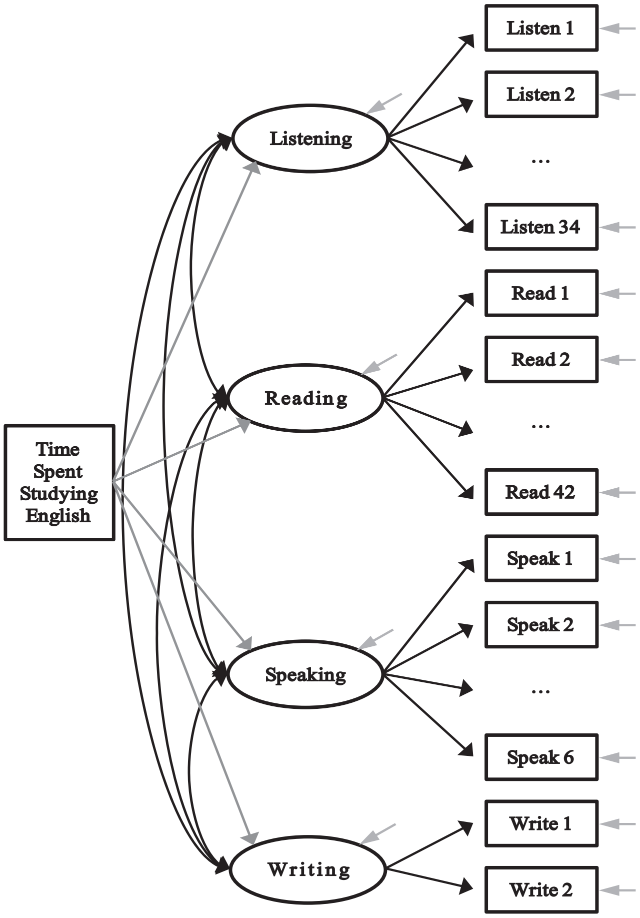

Figure 1 Multiple indicators and multiple causes (MIMIC) model for the covariate, time spent studying English.

English, there are four groups (g = 4); therefore, three (g −1) binary codes were created to identify three of the four levels of this covariate. In this case, the reference group (less than 2 years) did not receive its own dummy code, and mean of latent factor was specified as zero to allow mean difference comparisons with focal groups. The reference groups for the covariates used in this study are identified in Tables 2 and 3.

Before interpreting the effect of the covariate on the factors, the overall goodness-of-fit indices described above were examined to validate that the addition of the covariate did not change the factor structure of the baseline model and that the factor loadings, factor correlations, and residual variances/covariances were unchanged. In addition to significance, testing effect sizes were provided to guide interpretation of the effect of the covariate on the latent factor. However, caution should be used in interpretation, as these effect sizes do not imply causality, and there may be other factors not studied that impact the latent factor means. The effect size was calculated as square root of $[R^2/(1-R^2)]$, where R^2 is the percent of variance explained by the latent factors (Hancock, 2001). These effect sizes are similar to Cohen's (1988) F, where values of .1, .25, and .4 can be interpreted as small, medium, and large effect size, respectively.

Results

Confirmatory Factor Analysis

The overall goodness-of-fit statistics from the CFA of the six hypothesized models are presented in Table 4. In terms of choice of the baseline model, the model fit values for both the bifactor and correlated four-factor model all exceeded the criterion for each index: RMSEA values less than .05 and CFI and TLI values greater than .95. Moreover, the fit indices were very similar across the two models. An examination of the parameter estimates for the correlated four-factor model indicated that they were reasonable and statistically significant with completely standardized factor loading between .38

Table 4 Goodness-of-Fit Indices for Six Hypothesized Models

Model	χ^2	*df*	CFI	TLI	RMSEA (90% CI)
Bifactor	29,380	3,311	.98	.97	.019 (.018 – .019)
Correlated four factor	32,533	3,389	.97	.97	.019 (.019 – .020)
Single factor	81,381	3,402	.93	.93	.032 (.032 – .032)
Correlated two factor					
Read/listen vs. write/speak	48,027	3,396	.96	.96	.024 (.024 – .024)
Read/listen/write vs. speak	50,465	3,401	.96	.96	.025 (.025 – .025)
Higher-order factor	36,606	3,391	.97	.97	.021 (.021 – .021)

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean squared error of approximation; CI = confidence interval.

and .83 (Table 5). Also provided in Table 5 are standard measurement errors that represent the percent of variance in the indicators that is not explained by the target latent factor. Note, the writing section contains only two items; thus, the unstandardized factor loadings for these two items were set to be equal for model identification purposes. For the bifactor model, however, it was difficult to interpret the results, as many of the factor loadings were low and, in the case of listening, negative. As a result, the correlated four-factor model was used as the baseline for comparison of the other four competing models. It should be noted that the integrated tasks loaded on all target modalities and had nonsignificant loadings on the secondary target modalities (e.g., the listening factor loading for the integrated speaking item).

The global fit indices for four competing models indicated good fit to the data. One exception was the CFI and TLI values (.93) for the single-factor model that was slightly lower than .95. The single-factor model also had an RMSEA value higher than that of the correlated four-factor model. The correlated two-factor models had slightly lower CFI and TLI values (.96 and .96) compared with the correlated four-factor model (.97 and .97); the RMSEA values (.024 and .025) were also higher than the baseline correlated four-factor model. Although the higher-order factor model provided a fit to the data that was comparable with the correlated four-factor model (RMSEA = .021, CFI = .97, TLI = .97), an examination of the factor loadings and correlation between the factors suggested potential issues with model identification and interpretations. Specifically, low completely standardized factor loadings (<.10) were often found for the listening factor. In addition, the correlation between the general factor and the listening factor was .99, an indication that the two factors are not distinct from each other (Bagozzi & Yi, 1988).

The correlated four-factor model consistently provided a better fit to the data in terms of fit indices, as well as factor loading estimate and its significance level. In addition, the correlation pattern between the latent factors suggests that the abilities assessed by the four language components were distinct but moderately correlated (Table 6). The highest correlation was observed between the reading and listening factors (.89), whereas the reading and speaking factors had the lowest correlation (.63). Thus, the correlated four-factor model was chosen as the measurement model to use in the second set of analyses, CFA with covariates, in which the background variables were added independently to model the relationship between the background variables and the factors.

Confirmatory Factor Analysis With Covariates

Table 7 shows the overall fit of the correlated four-factor CFA model with the addition of the covariates: reason for taking the TOEFL iBT test, time spent studying English, time spent attending content classes taught in English, and lived in a country where English is the main language. Although sample size was reduced to exclude the nonresponses on the background questions, the MIMIC models all provided a good fit to the data with RMSEA between .017 – .018, and both CFI and TLI values between .97 – .98. Thus, the inclusion of the covariates did not alter the underlying correlated four-factor latent structure. The results also indicated that factor means were statistically significant at most levels of the covariates. The interpretations of model results are presented next.

Main Reason for Taking the TOEFL Test

Table 8 presents selected results from the MIMIC solution for the covariate, main reason for taking the TOEFL test. Note, the standardized parameter estimates of item factor loadings, not presented in Table 8, were the same or close to the values

Table 5 Standardized Parameter Estimates for the Correlated Four-Factor Model

Independent item	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)
1	.48 (.77)	.64 (.59)	.77 (.40)	.83 (.32)
2	.64 (.59)	.62 (.62)	.79 (.38)	
3	.44 (.81)	.74 (.45)		
4	.69 (.52)	.40 (.84)		
5	.42 (.83)	.50 (.75)		
6	.66 (.56)	.57 (.68)		
7	.77 (.40)	.56 (.68)		
8	.47 (.78)	.54 (.71)		
9	.52 (.73)	.57 (.67)		
10	.73 (.47)	.62 (.62)		
11	.58 (.66)	.46 (.79)		
12	.56 (.69)	.52 (.73)		
13	.63 (.61)	.61 (.62)		
14	.55 (.70)	.47 (.78)		
15	.55 (.75)	.57 (.67)		
16	.53 (.72)	.51 (.74)		
17	.50 (.75)	.51 (.74)		
18	.69 (.52)	.54 (.71)		
19	.59 (.65)	.64 (.60)		
20	.66 (.57)	.63 (.60)		
21	.56 (.68)	.38 (.86)		
22	.58 (.66)	.52 (.73)		
23	.39 (.85)	.64 (.59)		
24	.45 (.79)	.53 (.71)		
25	.66 (.56)	.54 (.71)		
26	.68 (.53)	.39 (.85)		
27	.39 (.85)	.41 (.84)		
28	.56 (.69)	.49 (.76)		
29	.55 (.70)	.46 (.79)		
30	.69 (.53)	.57 (.67)		
31	.49 (.76)	.59 (.65)		
32	.50 (.75)	.54 (.71)		
33	.53 (.72)	.74 (.45)		
34	.50 (.75)	.52 (.73)		
35	.48 (.77)			
36	.52 (.73)			
37	.45 (.80)			
38	.55 (.70)			
39	.61 (.63)			
40	.74 (.45)			
41	.53 (.72)			
42	.54 (.71)			
Integrated item	**Reading estimate (error)**	**Listening estimate (error)**	**Speaking estimate (error)**	**Writing estimate (error)**
1 (R/L/S)	.12	.03[a]	.73 (.31)	
2 (R/L/S)	.03[a]	.12	.72 (.31)	
3 (L/S)		−.01[a]	.83 (.33)	
4 (L/S)		.13	.72 (.33)	
5 (R/L/W)	.08	.20		.59 (.33)

[a] Nonsignificant ($|t| < 1.96; p > .05$); R/L/S = integrated speaking item with reading and listening; L/S = integrated speaking item with listening; R/L/W = integrated writing item with reading and listening.

shown in Table 5 for the baseline correlated four-factor model. Test takers whose intent was to attend graduate school served as the reference group in this MIMIC model. Statistically, significant differences were evident in the means for the reading, listening, speaking, and writing latent factors for test takers whose intent was graduate studies compared with those who took the TOEFL test for other reasons (p-values = .00). Given that the parameter estimates were all negative, it can be concluded that test takers whose intent was graduate studies always did better on the four latent language skills

Table 6 Correlation Across the Latent Factors for the Correlated Four-Factor Model

	Reading	Listening	Speaking	Writing
Reading	—			
Listening	.89	—		
Speaking	.63	.76	—	
Writing	.79	.77	.79	—

Table 7 Goodness-of-Fit Indices of Multiple Indicators and Multiple Causes (MIMIC) Models for Single Covariate

Background variable	*N*	χ^2	*df*	CFI	TLI	RMSEA (90% CI)
Reason for taking the TOEFL test	22,014	31,899	3,789	.97	.97	.018 (.018–.019)
Time spent studying English	11,373	15,873	3,629	.97	.97	.017 (.017–.017)
Time spent attending content classes taught in English	10,905	15,795	3,709	.98	.98	.017 (.017–.018)
Lived in a country where English is the main language	11,010	15,972	3,629	.98	.98	.018 (.017–.018)

Note: df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean squared error of approximation; CI = confidence interval. The total sample size of baseline model was 22,624 (CFI = .97; TLI = .97; RMSEA = .019).

Table 8 Summary of Multiple Indicators and Multiple Causes (MIMIC) Model Results: Reason for Taking the TOEFL Test

Factor	Group	Estimate (Std.)	*SE*	*p*-value	Effect size
Reading	High school/2-year college	−.69 (−.66)	.03	.00	.29
	Undergraduate	−.59 (−.57)	.02	.00	
	Licensure	−.20 (−.19)	.04	.00	
	Employment/immigration	−.19 (−.18)	.05	.00	
	Other	−.35 (−.34)	.03	.00	
Listening	High school/2-year college	−.61 (−.60)	.03	.00	.22
	Undergraduate	−.41 (−.40)	.02	.00	
	Licensure	−.14 (−.14)	.04	.00	
	Employment/immigration	−.14 (−.13)	.06	.01	
	Other	−.26 (−.26)	.03	.00	
Speaking	High school/2-year college	−.47 (−.46)	.03	.00	.13
	Undergraduate	−.15 (−.15)	.02	.00	
	Licensure	**.03 (.03)**	**.04**	**.44**	
	Employment/immigration	**−.07 (−.07)**	**.05**	**.21**	
	Other	−.09 (−.09)	.03	.00	
Writing	High school/2-year college	−.62 (−.61)	.03	.00	.19
	Undergraduate	−.30 (−.29)	.02	.00	
	Licensure	−.26 (−.26)	.05	.00	
	Employment/immigration	−.25 (−.24)	.05	.00	
	Other	−.28 (−.27)	.03	.00	

Note: Std. = standardized estimate; *SE* = standard error. Bolded values indicate nonsignificant ($p > .05$ estimates).

measured by the TOEFL test. One exception is on the speaking factor, where no difference was present between test takers whose intent was graduate studies versus those who took the test for licensure, employment, or immigration purposes (bolded estimates in Table 8). As an example of interpreting the standardized estimates presented, the estimate of −.66 for reading can be interpreted as indicating that those whose intent is to attend high school or 2-year college obtained .66 standardized score lower on the latent factor of reading than those whose intent was graduate school. Also provided in Table 8 are estimates of effect size. The effect size on each of the latent factors (reading = .29, listening = .22, speaking = .13, and writing = .19) indicated that, overall, the effect sizes on the listening, speaking, and writing latent factors among the test takers with various reason for taking the TOEFL test was small. However, there was a medium effect size on the reading latent factor.

Table 9 Summary of Multiple Indicators and Multiple Causes (MIMIC) Model Results: Language Exposure Covariates

Background question	Factor	Group	Estimate (Std.)	*SE*	*p*-value	Effect size
Time spent studying English	Reading	2 years or more, but less than 5 years	.25 (.24)	.04	.00	.30
		5 years or more, but less than 10 years	.61 (.59)	.04	.00	
		10 years or more	.94 (.90)	.04	.00	
	Listening	2 years or more, but less than 5 years	.35 (.33)	.04	.00	.31
		5 years or more, but less than 10 years	.65 (.62)	.04	.00	
		10 years or more	1.01 (.96)	.04	.00	
	Speaking	2 years or more, but less than 5 years	.31 (.30)	.05	.00	.33
		5 years or more, but less than 10 years	.62 (.59)	.04	.00	
		10 years or more	1.03 (.98)	.04	.00	
	Writing	2 years or more, but less than 5 years	.37 (.35)	.05	.00	.36
		5 years or more, but less than 10 years	.74 (.70)	.04	.00	
		10 years or more	1.15 (1.08)	.04	.00	
Time spent attending content classes taught in English	Reading	Less than 2 years	−.10 (−.10)	.03	.00	.13
		2 years or more, but less than 5 years	−.14 (−.14)	.03	.00	
		5 years or more, but less than 10 years	**−.03 (−.03)**	**.03**	**.34**	
		10 years or more	.24 (.24)	.03	.00	
	Listening	Less than 2 years	**−.02 (−.02)**	**.03**	**.50**	.16
		2 years or more, but less than 5 years	.07 (.07)	.03	.02	
		5 years or more, but less than 10 years	.15 (.15)	.03	.00	
		10 years or more	.43 (.43)	.03	.00	
	Speaking	Less than 2 years	.08 (.08)	.03	.01	.28
		2 years or more, but less than 5 years	.23 (.23)	.03	.00	
		5 years or more, but less than 10 years	.38 (.36)	.03	.00	
		10 years or more	.82 (.79)	.03	.00	
	Writing	Less than 2 years	.10 (.09)	.03	.00	.21
		2 years or more, but less than 5 years	.16 (.16)	.03	.00	
		5 years or more, but less than 10 years	.31 (.30)	.04	.00	
		10 years or more	.61 (.60)	.04	.00	
Lived in a country where English is the main language	Reading	Less than 1 year	−.05 (−.05)	.03	.04	.08
		1 year or more, but less than 5 years	**−.06 (−.06)**	**.03**	**.06**	
		5 years or more	.28 (.28)	.04	.00	
	Listening	Less than 1 year	.07 (.07)	.03	.01	.12
		1 year or more, but less than 5 years	.18 (.18)	.03	.00	
		5 years or more	.46 (.45)	.04	.00	
	Speaking	Less than 1 year	.18 (.17)	.03	.00	.24
		1 year or more, but less than 5 years	.37 (.36)	.03	.00	
		5 years or more	.90 (.88)	.04	.00	
	Writing	Less than 1 year	.07 (.07)	.03	.01	.14
		1 year or more, but less than 5 years	.20 (.20)	.03	.00	
		5 years or more	.55 (.54)	.05	.00	

Note: Std. = standardized estimate; *SE* = standard error. Bolded values indicate nonsignificant ($p > .05$ estimates).

Covariates Related to Language Exposure

Table 9 presents the MIMIC results for the background variables measuring exposure to English. These include time spent studying English, time spent attending content class taught in English, and lived in a country where English is the main language. As previously, although not presented, the estimates for item factor loadings were the same or close to the values for the baseline model presented in Table 5. The standardized estimates are also diagramed in Figure 2 for ease of comparison. Specifically, each graph shows that standardized estimate at the different level of the covariate measuring exposure to the target language in comparison with the reference group.

For the first covariate, time spent studying English, significant differences (*p*-values = .00) were observed in the means of the latent factors (reading, listening, speaking, and writing) between those who spent less than 2 years studying English (reference group) and the other three levels of time spent studying English (comparison groups). In addition, the parameter estimates were all positive; that is, as time spent studying English increased, the mean on the latent factors increased. For example, the standardized score on reading was .24 units higher for those who spent at least 2 years but less than

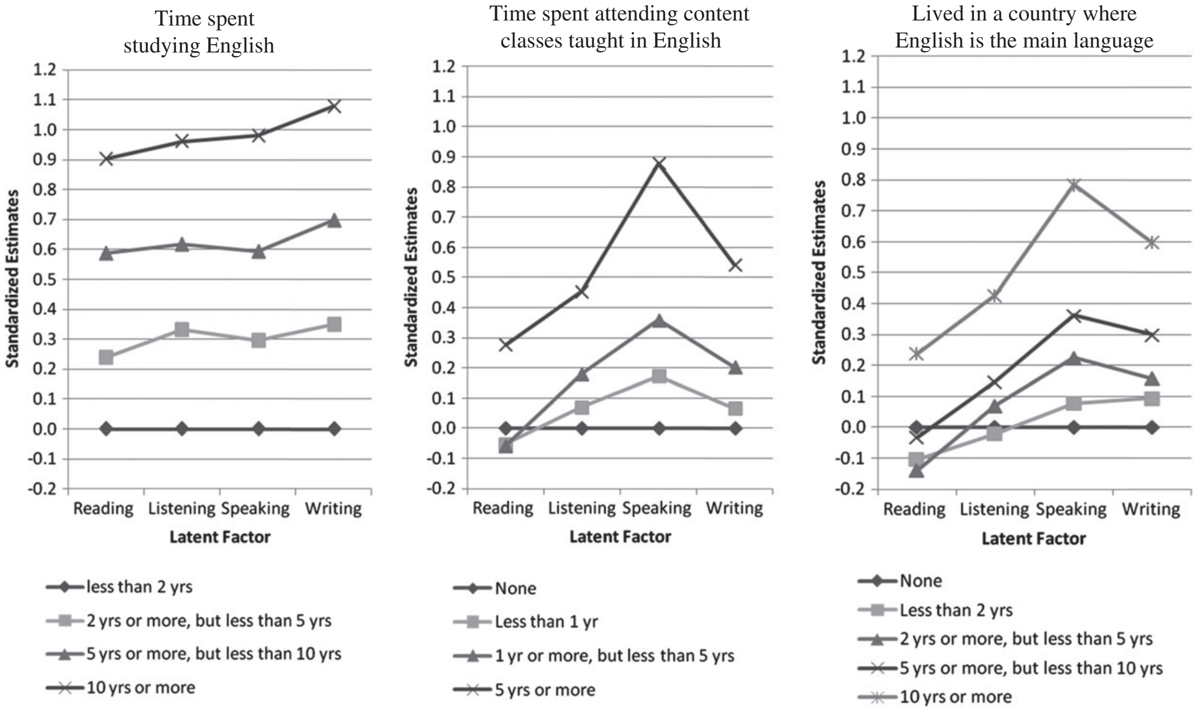

Figure 2 Standardized parameter estimates for language exposure covariates.

5 years studying English compared with those who spent less than 2 years, .59 units higher for those who studied at least 5 years but less than 10 years, and .90 units for those who studied for more than 10 years. Furthermore, the effect size on each of the latent factors (reading = .30, listening = .31, speaking = .33, and writing = .36) indicated meaningful differences among the groups relative to time spent studying English.

For the second covariate, time spent attending content classes taught in English, the reference group for the comparisons was test takers who had never attended content classes taught in English. The results indicated no practical significance on the reading and listening factors for those who attended content classes taught in English compared with those who never attended (effect size of .13 and .16, respectively). In fact, the parameter estimates at the first three levels of the covariate on the reading factor were negative (and nonsignificant for the level 5 years or more but less than 10 years). Nonsignificant differences were also observed on the listening factor for those who attended content classes taught in English for less than 2 years. In contrast, the impact of having attended content classes taught in English on the means speaking and writing factors were of practical significance with effect sizes of .28 and .21, respectively. In addition, the parameter estimates for the speaking and writing factors were significant and increased as the amount of time spent attending classes taught in English increased.

In modeling the third background question measuring exposure to English, those test takers who had not lived in a country where English was the main language served as the reference group. For the reading factor, a positive and significant parameter estimate was observed only for those test takers who lived in a country where English was the main language for more than 5 years. Moreover, the effect size of this covariate on the reading latent factor was negligible (.08). In contrast, significant differences on the listening, speaking, and writing factor means were observed at all levels. However, the effect size among the different levels of time lived in a country where English was the main language was small, with the exception of speaking, with an effect size of .24.

Interaction Between Language Family and Language-Exposure Covariates

The interaction effect of language family membership and language exposure covariates on the reading, listening, speaking, and writing latent factors are summarized in Tables 10 and 11. The standardized estimates in Table 11 are also diagramed in Figure 3. Table 10 shows the overall goodness-of-fit statistics for the three CFA models studied. All three

Table 10 Goodness-of-Fit Indices of Multiple Indicators and Multiple Causes (MIMIC) Models for Multiple Covariates

Background variable	*N*	χ^2	*df*	CFI	TLI	RMSEA (90% CI)
Language family membership and time spent studying English	11,218	15,994	3,629	.97	.97	.017 (.017–.018)
Language family membership and time spent attending content classes taught in English	10,756	17,442	3,629	.97	.97	.019 (.019–.019)
Language family membership and lived in a country where English is the main language	10,864	16,941	3,629	.97	.97	.018 (.018–.019)

Note: df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean squared error of approximation; CI = confidence interval. The total sample size of baseline model was 22,624 (CFI = .97; TLI = .97; RMSEA = .019).

Table 11 Summary of Multiple Indicators and Multiple Causes (MIMIC) Model Results: Language Family and Language Exposure

Background question	Factor	Group	Estimate (Std.)	*SE*	*p*-value	Effect size
Time spent studying English	Reading	NIE and 5 years or more	.61 (.57)	.03	.00	.37
		IE and less than 5 years	.52 (.49)	.04	.00	
		IE and 5 years or more	1.11 (1.04)	.03	.00	
	Listening	NIE and 5 years or more	.52 (.49)	.03	.00	.39
		IE and less than 5 years	.45 (.42)	.04	.00	
		IE and 5 years or more	1.14 (1.06)	.03	.00	
	Speaking	NIE and 5 years or more	.47 (.43)	.03	.00	.44
		IE and less than 5 years	.43 (.40)	.04	.00	
		IE and 5 years or more	1.23 (1.13)	.04	.00	
	Writing	NIE and 5 years or more	.60 (.57)	.03	.00	.33
		IE and less than 5 years	.13 (.12)	.04	.00	
		IE and 5 years or more	.92 (.88)	.04	.00	
Time spent attending content classes taught in English	Reading	NIE and 5 years or more	.24 (.23)	.03	.00	.27
		IE and less than 5 years	.57 (.55)	.03	.00	
		IE and 5 years or more	.66 (.64)	.03	.00	
	Listening	NIE and 5 years or more	.37 (.36)	.03	.00	.32
		IE and less than 5 years	.65 (.62)	.03	.00	
		IE and 5 years or more	.81 (.77)	.03	.00	
	Speaking	NIE and 5 years or more	.62 (.57)	.03	.00	.43
		IE and less than 5 years	.79 (.72)	.03	.00	
		IE and 5 years or more	1.18 (1.08)	.03	.00	
	Writing	NIE and 5 years or more	.41 (.40)	.03	.00	.23
		IE and less than 5 years	.34 (.33)	.03	.00	
		IE and 5 years or more	.67 (.65)	.04	.00	
Lived in a country where English is the main language	Reading	NIE and 1 year or more	.13 (.13)	.03	.00	.26
		IE and less than 1 year	.54 (.53)	.02	.00	
		IE and 1 year or more	.62 (.60)	.04	.00	
	Listening	NIE and 1 year or more	.40 (.38)	.03	.00	.31
		IE and less than 1 year	.65 (.62)	.03	.00	
		IE and 1 year or more	.80 (.77)	.04	.00	
	Speaking	NIE and 1 year or more	.61 (.57)	.03	.00	.42
		IE and less than 1 year	.78 (.72)	.02	.00	
		IE and 1 year or more	1.29 (1.19)	.04	.00	
	Writing	NIE and 1 year or more	.33 (.33)	.04	.00	.20
		IE and less than 1 year	.34 (.33)	.03	.00	
		IE and 1 year or more	.62 (.61)	.04	.00	

Note: Std. = standardized estimate; *SE* = standard error; IE = Indo-European; NIE = non-Indo-European.

models had acceptable fit statistics (RMSEA between .017 and .019, and CFI and TLI values of .97), indicating that the addition of the covariates did not negatively impact the fit of the correlated four-factor model to the data. As seen in Table 11, significant differences were observed in the means of the latent factors, depending on language family and covariates measuring language exposure at all levels. Furthermore, the parameter estimates measuring the effect of these variables on the latent factors had an overall effect size that ranged from medium to large at most levels. In particular, the effect size on the speaking factor for all three models was larger than .40.

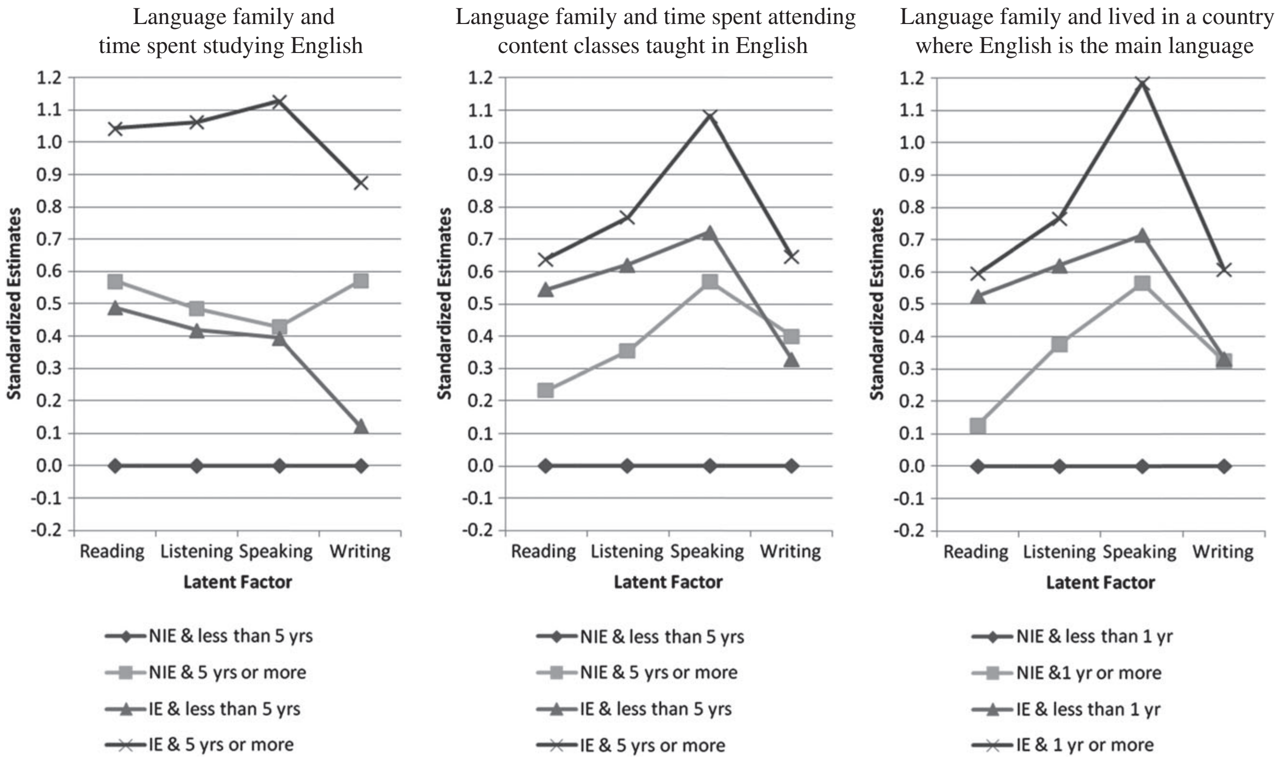

Figure 3 Standardized parameter estimates for language family and language exposure covariates.

When grouped by language family membership and by time spent studying English, the results indicated that the IE test takers always had higher mean estimates on the four latent factors than NIE test takers for the same amount of time spent studying English. For example, in comparison with the NIE and less-than-5-years-studying-English test takers (reference group), the NIE group with more than 5 years studying English was .43 standardized units higher on the speaking factor, whereas the IE group that studied English for more than 5 years was 1.3 units higher on the same factor.

When language exposure was measured by test takers who had attended classes taught in English, the highest mean on the latent factors was obtained by the IE test takers with more than 5 years of attending classes in comparison with the reference group (i.e., NIE test takers with less than 5 years of attending classes taught in English). In contrast, with the exception for the writing factor, NIE test takers who attended classes taught in English more than 5 years had lower latent factor means on reading, listening, and speaking.

Similar to the two previous analyses, test takers in the IE language family for the highest level of exposure to English, as measured by having lived in a country where English is the main language for more than 1 year, had the highest mean on the four latent factors. This result was followed by the IE language group with less than 1 year of having lived in a country where English is the main language. The NIE and more-than-1-year group had the lowest means on the reading, listening, and speaking factor. The mean on the writing factor was similar to the IE and less-than-1-year group.

Discussion

The purpose of this study was to determine the relationship between test-taker background characteristics and performance on the underlying ELP traits as measured by the TOEFL iBT test. To address the research question, a MIMIC approach was used to model the direct effect of four test-taker background characteristics on the underlying latent structures that represent the TOEFL iBT test. As a first step, six competing factor structures (bifactor, correlated four-factor, single-factor, two versions of a correlated two-factor model, higher-order factor) were compared to determine the measurement model that best represented the TOEFL iBT test. This comparison showed that the correlated four-factor model representing the four language modalities of reading, listening, speaking, and writing best accounted for the underlying latent factor structure of the TOEFL iBT test. This result is consistent with a recent investigation by Sawaki and Sinharay (2013) into the factor structure of the TOEFL iBT test. However, these findings are different from that of Sawaki et al. (2008), who concluded that the higher-order factor model with a higher-order general factor representing ESL/EFL and four first-order factors for reading, listening, speaking, and writing represented the factor structure of the TOEFL iBT test. One possible explanation for this divergent finding was that Sawaki et al. used data from a field trial for the TOEFL iBT test, which may not be totally representative of the current TOEFL iBT test population. In contrast, Sawaki and Sinharay and this study used data from operational administrations of the TOEFL iBT test. Specifically, Sawaki and Sinharay used data for operational test forms administered in 2007, and this study used data from a more recent operational administration, Fall 2013.

In the next phase of the analyses, the background characteristics were added independently to the correlated four-factor model. The study results indicated that the MIMIC model provided a good fit to the data and that the addition of the covariates to the CFA model did not alter the underlying correlated four-factor model identified in the precursor to the covariate analyses. Moreover, at most levels of the covariates studied, there was a statistically significant effect on the latent factors; that is, the study results provided empirical support for the association and possible influence of the studied test-taker background characteristics on test performance.

The study results provide evidence that differences in performance on English-language tests such as TOEFL iBT are associated with intended test score use and associated stakes. When test takers were classified by reason for taking the TOEFL iBT test with test takers whose intent was graduate studies as the reference group, large differences were observed on the reading, listening, and writing factors, but less so on the speaking factor. This finding may reflect the more formal instruction in English at the higher educational level on the reading, listening, and writing factors and the higher English language demand expected in graduate studies. For example, the study results show test takers with less than 4 years of college (i.e., those whose intent was to attend high school or 2-year colleges) had lower means on the four latent factors in comparison with the reference group. In addition, no significant differences were observed on the speaking latent means for potential graduate students and those who took the TOEFL iBT test for immigration and employment purposes. This result probably reflected the similar and higher speaking proficiency requirements for these purposes. The finding is supported by the relatively high TOEFL iBT Speaking score requirement of 26 (0–30 scale) by health care professional associations and the importance of communicative ability in providing quality counseling and patient care (National Association of Boards of Pharmacy, 2012; Wendt & Woo, 2009).

In addition, the results of this study support the findings of previous studies, which have shown that the greater familiarity and exposure formally or informally to English, the greater the ELP of test takers will be (Bachman, 1990; Gradman & Hanania, 1991; Kunnan, 1995). The findings provided convincing support for the claim that the greater the exposure to English, the higher the means on the latent language abilities, as measured by the TOEFL iBT test. There were, however, differential impacts on the latent components, depending on the exposure covariate. For example, when compared to test takers who spent less than 2 years studying English, substantial differences were observed in the means of the latent factors of reading, listening, speaking, and writing at increased levels of time spent studying English. The improvements were somewhat consistent across the four language skills. In contrast, test takers who attended a school, college, or university in which content classes were taught in English performed better on speaking compared with those who did not have this experience. Smaller positive differences were observed on the writing and listening factors. Interestingly, this covariate had a positive impact on the reading factor only when the test taker attended content classes taught in English for more than 10 years. The effect of having lived in a country where English is the main language also had a more pronounced positive effect on the mean of the latent factor measuring speaking and smaller positive effects were observed on the reading, listening, and writing means. The pattern of results for these two covariates, attended content classes taught in English and lived in a country where English is the main language, indicate support of the hypothesis that context and instruction or exposure to speakers of English is an important factor contributing to language learning.

Further, when different language family groups were compared within the three background characteristics measuring prior exposure to English, the IE language group showed higher means of the latent factors than the NIE group. The results of this study also confirm the hypothesis that background characteristics measuring exposure to English have a greater positive impact on test takers whose native language is in the same language family as English. Specifically, the IE group always performed better than the NIE group when they experienced the same level of prior exposure to English. The largest impact was observed on the speaking factor.

In summary, the central finding from this study was the identification of test-taker background characteristics associated with differences in English-language test performance. However, the purpose of this study was to measure the association and possible influence of test-taker background characteristics on test performance in the context of the TOEFL iBT test, which is mostly taken by adult English-language learners; thus, the generalizability of the study findings was limited. Therefore, it is important to replicate this study using other language tests, as well as different background variables and categorization representing exposure to the target language. Moreover, no information was available on the age at which the test takers experience the specific English-language exposures in this study. Future studies should consider this factor, as it might explain observations such as less influential impact on the reading factor found in this study.

References

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

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*. New York, NY: Cambridge University Press.

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

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.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*, 74–94.

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

Carroll, J. B. (1965). Fundamental consideration in testing for English language proficiency of foreign students. In H. B. Allen (Ed.), *Teaching English as a second language: A book of readings* (pp. 364–372). New York, NY: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Educational Testing Service. (2013). *Test and score data summary for TOEFL iBT test January 2013–December 2013 test data*. Princeton, NJ: Author.

Fabriger, 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*, 272–299.

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.

Gradman, H., & Hanania, E. (1991). Language learning background factors and ESL proficiency. *Modern Language Journal*, *75*, 39–51.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, *31*, 111–133.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373–388.

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*, 1–55.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing*, *29*, 131–152.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Kachru, B. (1984). World Englishes and the teaching of English to non-native speakers: Context, attitudes, and concerns. *TESOL Newsletter*, *18*, 25–26.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 23–44). Westport, CT: American Council on Education and Praeger.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural equation modeling approach*. Cambridge, England: Cambridge University Press.

Kunnan, A. J. (1998). Approach to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 1–16). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Authors.

National Association of Boards of Pharmacy. (2012). *Foreign pharmacy graduate examination committee application bulletin*. Retrieved from http://www.nabp.net/programs/assets/FPGECBulletin.pdf

Oller, J. W., Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen*, *76*, 165–174.

Oller, J. W., Jr. (1979). The factorial structure of language proficiency: Divisible or not? In J. W. Oller Jr. (Ed.), *Language test at school: A pragmatic approach* (pp. 423–458). London, England: Longman.

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

Powers, D. E. (2010). *The case for a comprehensive, four-skills assessment of English language proficiency* (TOEIC Compendium Study TC-10-12). Princeton, NJ: Educational Testing Service.

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. TOEFLiBT-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02342.x

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Report No. TOEFLiBT-04). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02095.x

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

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (TOEFL iBT Research Report No. TOEFLiBT-07). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02152.x

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Report No. TOEFL-RR-06). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1980.tb01229.x

Wendt, A., & Woo, A. (2009, August). *A minimum English proficiency standard for the Test of English as a Foreign Language Internet-based test* (NCLEX® Psychometric Research Brief). Retrieved from https://www.ncsbn.org/TOEFL_iBT_Proficiency_Standard_Process.pdf

Wilson, K. M. (2000). *An exploratory dimensionality assessment of the TOEIC test* (Research Report No. RR-00-14). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2000.tb01837.x

Wolf, M. K., Kao, J., Herman, J., Bachman, L. F., Bailey, A., Bachman, P. L., … Chang, S. M. (2008). *Issues in assessing English language learners*: *English language proficiency measures and accommodation uses—Literature review* (*Part 1 of 3*) (*CRESST Report No*. *731*). Los Angeles, CA: CRESST/UCLA.

Appendix

Diagrams of Six Confirmatory Factor Analysis (CFA) Models

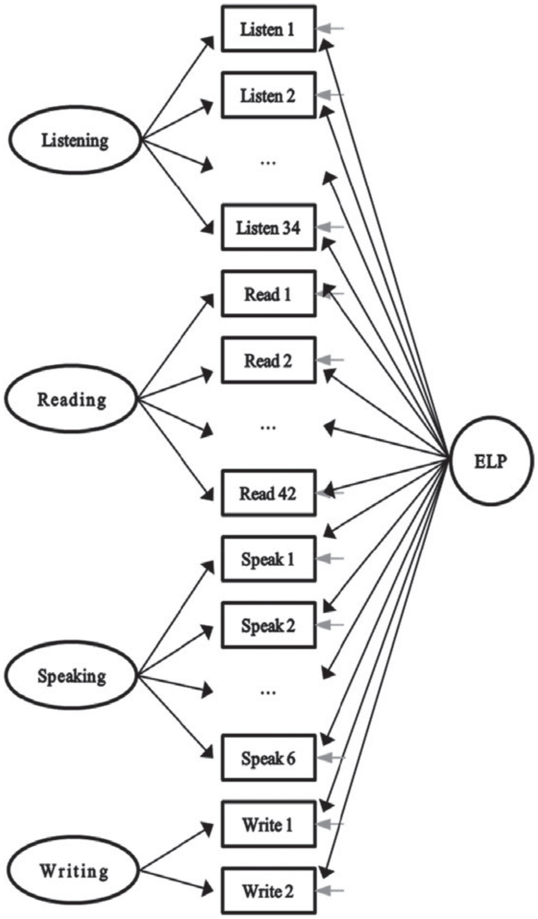

Figure A1 Bifactor model.

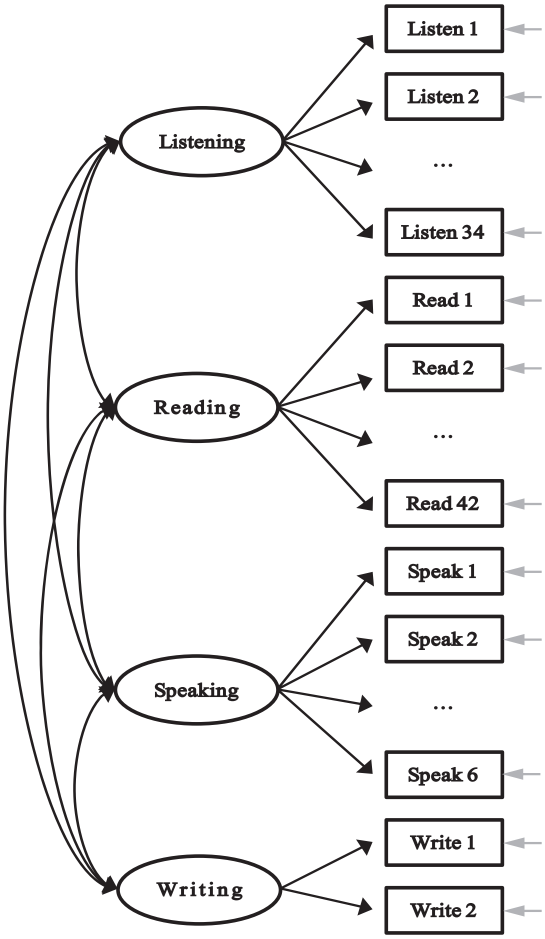

Figure A2 Correlated four-factor model.

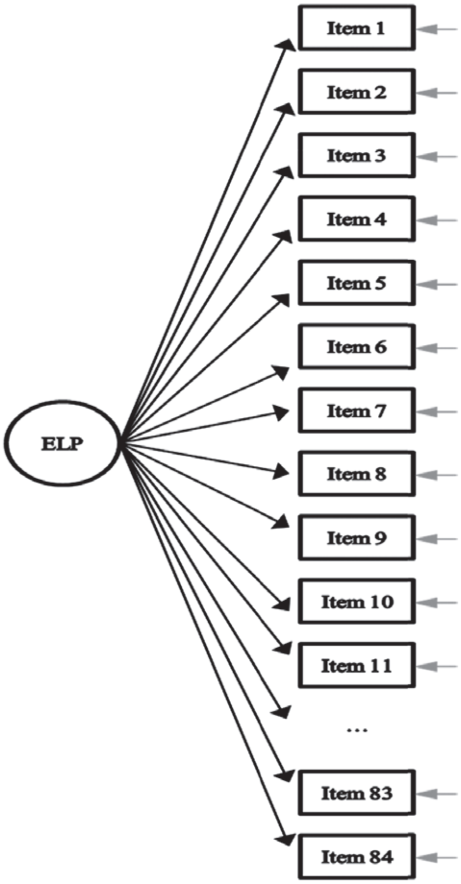

Figure A3 Single-factor model.

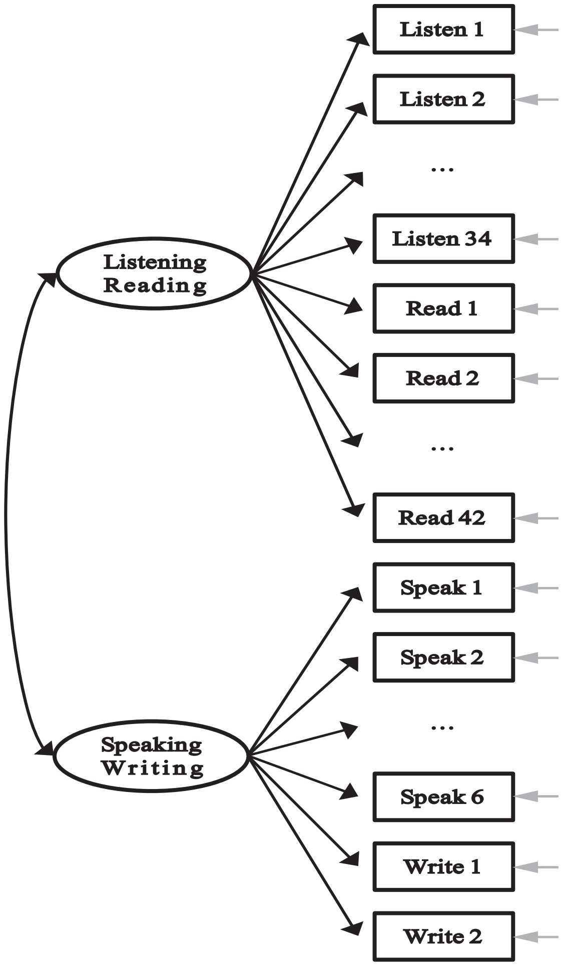

Figure A4 Correlated two-factor model.

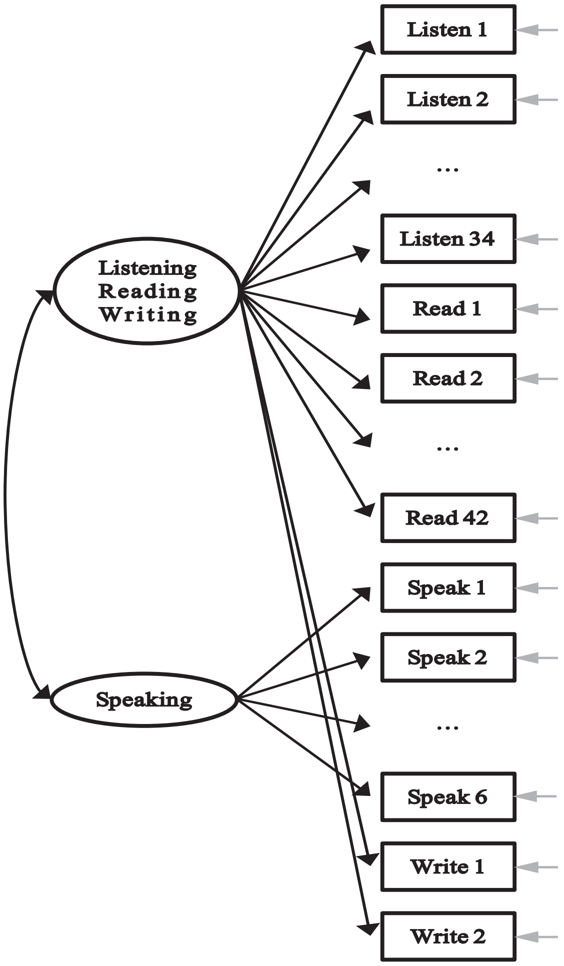

Figure A5 Correlated two-factor model.

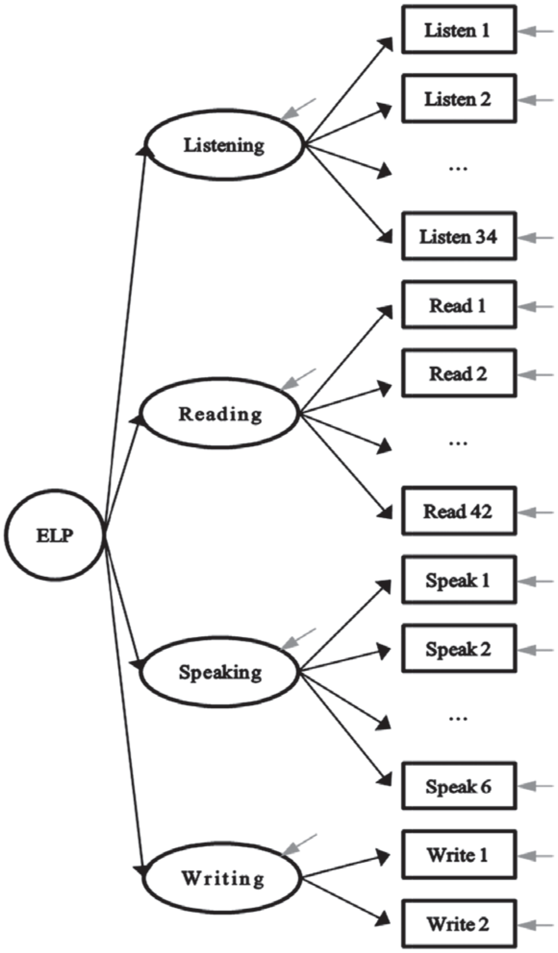

Figure A6 Higher-order factor model.

Suggested citation:

Manna, V. F., & Yoo, H. (2015). *Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English-as-a-second-language (ESL) test population: A factor analytic approach* (Research Report No. RR-15-25). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12072

Action Editor: Don Powers

Reviewers: Mikyung Wolf and Lixiong Gu