ISSN 1930-9317

TOEFL

TOEFL iBT Research Report

TOEFLiBT-07
December 2008

Factor Structure of the TOEFL Internet-Based Test Across Subgroups

Lawrence J. Stricker

Donald A. Rock

Listening.
Learning.
Leading.®

Factor Structure of the TOEFL® Internet-Based Test Across Subgroups

Lawrence J. Stricker and Donald A. Rock

ETS, Princeton, NJ

RR-08-66

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Copyright © 2008 by ETS. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service (ETS). The TEST OF ENGLISH AS A FOREIGN LANGUAGE is a trademark of ETS.

College Board is a registered trademark of the College Entrance Examination Board.

Abstract

This study assessed the invariance in the factor structure of the *Test of English as a Foreign Language*™ Internet-based test (TOEFL® iBT) across subgroups of test takers who differed in native language and exposure to the English language. The subgroups were defined by (a) Indo-European and Non-Indo-European language family, (b) Kachru's classification of outer and expanding circles of countries (based on prevalence of English use in educational and business contexts), and (c) years of classroom instruction in the English language. The same factor structure (four first-order factors corresponding to the test sections and a single higher-order factor encompassing these factors) was identified in each subgroup. The results support the present scoring scheme for the TOEFL iBT assessment and suggest that the test functions the same way for diverse subgroups of test takers.

Key words: English-language study, factor analysis, Indo-European, Kachru, subgroups of test takers, TOEFL iBT

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2007-2008) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Geoffrey Brindley	Macquarie University
Frances A. Butler	Language Testing Consultant
Carol A. Chapelle	Iowa State University
Catherine Elder	University of Melbourne
April Ginther	Purdue University
John Hedgcock	Monterey Institute of International Studies
David Mendelsohn	York University
Pauline Rea-Dickins	University of Bristol
Mikyuki Sasaki	Nagoya Gakuin University
Steven Shaw	University of Buffalo

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

Thanks are due to Jill Carey and Lin Wang for providing information about the TOEFL iBT field test; Min Hwei Wang for doing the computer analysis; and Lydia Liu, Frank Rijmen, and Todd Rogers for reviewing a draft of this report.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Important evidence is beginning to accumulate about the construct validity of the *Test of English as a Foreign Language*™ Internet-based test (TOEFL® iBT; see the review by Chappelle, Enright, & Jamieson, 2008). Particularly relevant is a recent confirmatory factor analysis (Sawaki, Stricker, & Oranje, 2008) that found that the test assessed four first-order factors corresponding to the four sections for the test (Listening, Reading, Speaking, Writing) and a single higher-order factor that encompassed the first-order factors. This result is consistent with the consensus in the language-testing literature that language ability is multicomponential, with a higher-order, general factor as well as with smaller group factors (e.g., see the review by Sasaki, 1999). This outcome also supports the policy of reporting four scores for the test: one for each section and a single composite score.

An unresolved question is whether the same result, based on an aggregate sample of test takers in a field study, drawn from 93 home countries and differing greatly in their backgrounds, would be observed in relatively homogeneous subgroups of test takers varying in their native language and exposure to English. The TOEFL iBT assessment might be expected to be sensitive to these differences in so far as they affect the acquisition of English, and such effects should be evident in subgroup differences in the test's factor structure. In the case of formal exposure to the English language, for example, if English-language training emphasized reading, writing, and listening, with little attention given to speaking, then reading, writing, and listening factors might be highly correlated with each other and minimally correlated with the speaking factor (Stricker, Rock, & Lee, 2005).

Test takers' language family, particularly Indo-European and Non-Indo-European, has been widely investigated in language-testing research (e.g., Kunnan, 1995) and has been examined in a study of the TOEFL iBT assessment that is now underway (Xi, Midouhas, & Steinberg, 2006). Prevalence of English use in educational and business contexts in the test takers' home country, particularly Kachru's (1984; 1985) classification of inner-circle countries (English is primary; e.g., United States), outer-circle countries (English has special administrative status; e.g., India), and expanding-circle countries (English is considered important but has no special administrative status; e.g., Japan), is frequently used in the English as a second language/English as a foreign language teaching and learning literature (e.g., Thumboo, 2001) and was recently investigated in a TOEFL iBT study (Xi et al., 2006). Test takers' language exposure, especially in a formal school

setting, has been investigated in language-testing research (e.g., Kunnan, 1995) and in a recent TOEFL iBT study (Xi et al., 2006).

Accordingly, the aim of the present study was to assess the invariance of the factors underlying the TOEFL iBT assessment for each of the three kinds of subgroups defined by (a) Indo-European and Non-Indo-European language family, (b) Kachru's (1984; 1985) outer- and expanding-circle countries, and (c) years of classroom instruction in the English language.

Method

Sample

The samples were drawn from the 2,720 test takers, paid participants recruited and tested in 30 countries, intended to approximate the TOEFL test-taking population, who took the TOEFL iBT assessment in a 2003–2004 field study (Wang, Eignor, & Enright, 2008). The same test takers were used in the previous factor analysis of the total sample (Sawaki et al., 2008). (Subgroup analyses in that study were precluded because the sample sizes were inadequate for the large number of variables involved in the item-level analyses that were conducted.) In defining the subgroups for the present study, test takers who took the TOEFL iBT assessment at a test center in Australia, Canada, Great Britain, or the United States were excluded to minimize the irrelevant effects of incidental exposure to English. The three sets of samples were:

1. Language family: Indo-European ($N = 657$) and Non-Indo-European ($N = 669$).[1]
2. Kachru's (1984, 1985) inner-outer-expanding-circle classification of native country: outer-circle countries ($N = 311$) and expanding-circle countries ($N = 379$).[2]
3. Amount of classroom English-language study: 6 years or less ($N = 585$), 7 to 10 years ($N = 407$), and 11 years or more ($N = 406$).[3]

The TOEFL iBT scores of the samples are summarized in Table 1. (Scaled scores that range from 0 to 30 are reported for each section; the total score is the sum of the scaled scores for the four sections.) Within the three sets of samples, the mean section and total scores were generally similar. Consistent exceptions were the Listening, Speaking, Writing, and total means for 6 years or less and 7 to 10 years of English-language study. The means for the latter sample were appreciably higher.

Table 1

TOEFL iBT Scores of the Samples

Sample	*N*	Listening		Reading		Speaking		Writing		Total	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Language family											
Indo-European	657	18.13	6.58	17.88	6.64	18.80	6.57	17.22	6.53	72.03	23.37
Non-Indo-European	669	16.01	6.91	16.18	6.85	15.60	7.36	14.90	6.81	62.68	25.06
Outer- and expanding-circle countries											
Outer circle	311	17.70	6.40	17.73	6.81	18.84	6.93	17.72	7.48	71.99	24.58
Expanding circle	379	16.20	7.00	16.88	6.95	15.37	7.44	14.90	6.59	63.35	25.27
English-language study											
6 years or less	585	14.62	6.71	14.82	6.53	14.27	6.89	13.45	5.81	57.15	22.77
7 to 10 years	407	18.15	6.58	17.86	6.55	17.88	6.57	16.72	6.45	70.61	23.30
11 years or more	406	19.40	6.19	19.31	6.50	20.38	6.55	18.78	7.02	77.88	23.44

Measures

A total of 17 scores from the Listening, Reading, Speaking, and Writing sections of the TOEFL iBT assessment were used in the analysis.

The Listening section consists of six prompts, each with five or six items (a total of 31 multiple-choice items scored dichotomously and two complex selected-response items scored polytomously, 0 to 2). A total score over the six prompts, converted to a scaled score, is reported for the test. For this study, a total score for the set of items for each prompt was obtained, and these six scores were used in the analysis. (Using the total score for a prompt, instead of using individual items, eliminates the experimental dependence among items associated with a particular prompt, as well as the instability inherent in factor analyses of items [Gorsuch, 1974], and reduces the sample size needed for the analysis.)

The Reading section consists of three passages, each with 12 to14 items (a total of 35 multiple-choice items scored dichotomously, and three complex selected-response items scored polytomously: 0 to 2, 0 to 3, or 0 to 4). A total score over the three passages, converted to a scaled score, is reported for the test. For this study, a total score for the set of items for each prompt was obtained, and these three scores were used in the analysis. The total score for a prompt was used for the reasons already described.

The Speaking section consists of six speaking tasks, four of which are integrated tasks measuring more than one skill: two for listening/speaking and two for reading/listening/speaking. Each task is rated on a 0 to 4 scale by experienced raters. The mean of these scores, converted to a scaled score, is reported for the test. For this study, the score for each task was obtained, and these six scores were used in the analysis.

The Writing section consists of two writing tasks, one of which is an integrated task: reading/listening/writing. Each task is rated on a scale of 0 to 5. The mean of these scores, converted to a scaled score, is reported for the test. For this study, the score for each task was obtained, and these two scores were used in the analysis.

Analysis

Multiple-group confirmatory factor analyses of the 17 scores for each of the three sets of samples were conducted. Scores for each sample were normalized, using an area conversion of scores, with PRELIS 2 (Joreskog & Sorbom, 1996b). Covariance matrices for each sample were

computed from the normalized scores and then analyzed by the robust maximum likelihood method with LISREL 8.53 (Joreskog & Sorbom, 1996a).

Hypotheses about the factors and their invariance were tested in two stages. First, competing, nested models were tested about the number of factors in the test. These models were based on viable factor solutions in previous confirmatory factor analyses of the TOEFL iBT assessment (Sawaki et al., 2008) and a prototype of this test, LanguEdge (ETS, 2002, 2004; Stricker et al., 2005). The models follow:

1. There is only one factor, made up of the four sections of the test (see Figure 1). This is an obvious model for cognitive tests and was a possible factor solution in the Stricker et al. (2005) study.
2. There are two correlated factors, one for the Speaking section and one for the Listening, Reading, and Writing sections (see Figure 2). This model, the final factor solution identified in the Stricker et al. (2005) study, was based on an exploratory factor analysis of a separate sample in that investigation.
3. There are four correlated first-order factors corresponding to the test sections. (see Figure 3). This model partially reflects the rationale underlying the test and was a possible factor solution in the Sawaki et al. (2008) study.
4. There are four correlated first-order factors corresponding to the test sections, and they are subsumed by a higher-order general factor (see Figure 4). This model completely reflects the rationale underlying the test and was the final solution identified in the Sawaki et al. (2008) study.

Second, based on the factor model that was best supported, hierarchically ordered nested models were tested about the invariance of the factors across samples. The nested models for Models 1, 2, and 3 were:

1. The number of factors is invariant.
2. The factor loadings are invariant.
3. The factor loadings and error variances are invariant.
4. The factor loadings, error variances, and intercorrelations are invariant. (This nested model is not relevant to Model 1.)

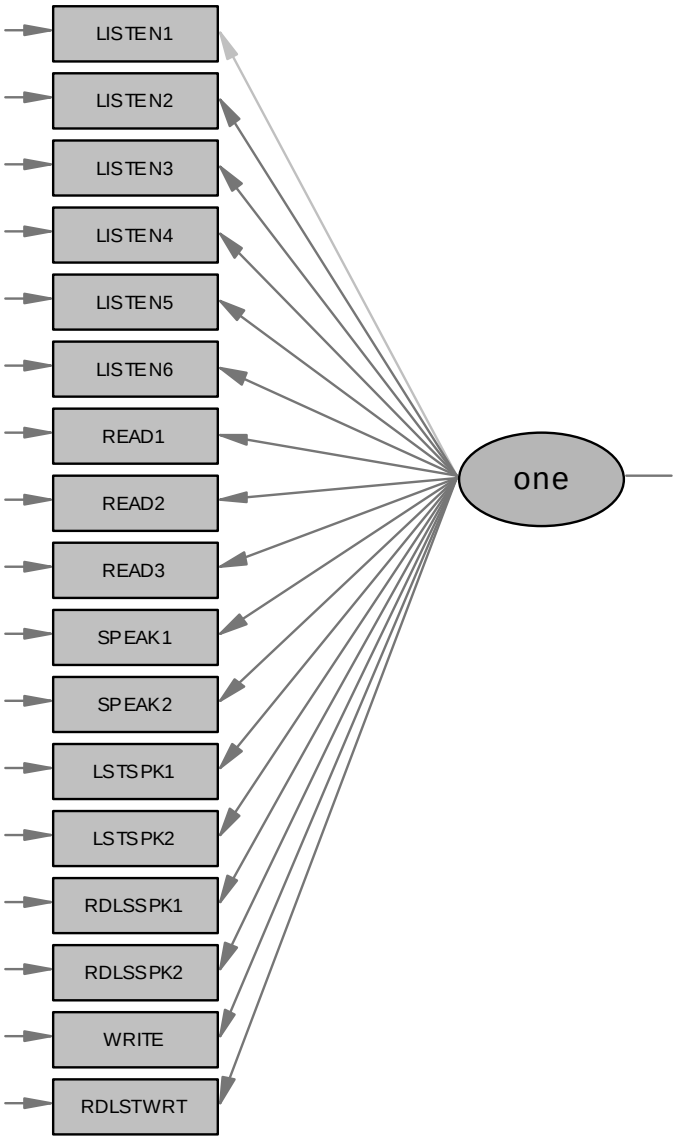

Figure 1. **Model 1: One factor—Listening, Reading, Speaking, and Writing.**

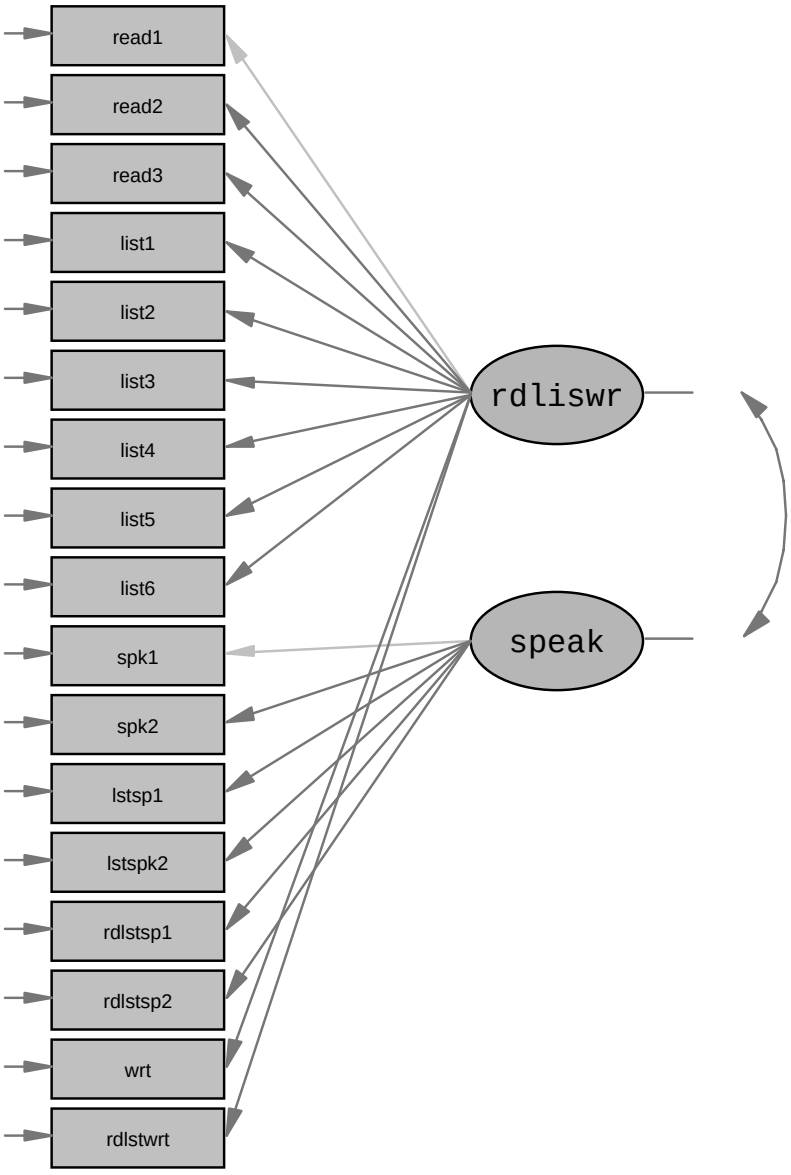

Figure 2. **Model 2: Two factors─Speaking vs. Listening, Reading, and Writing.**

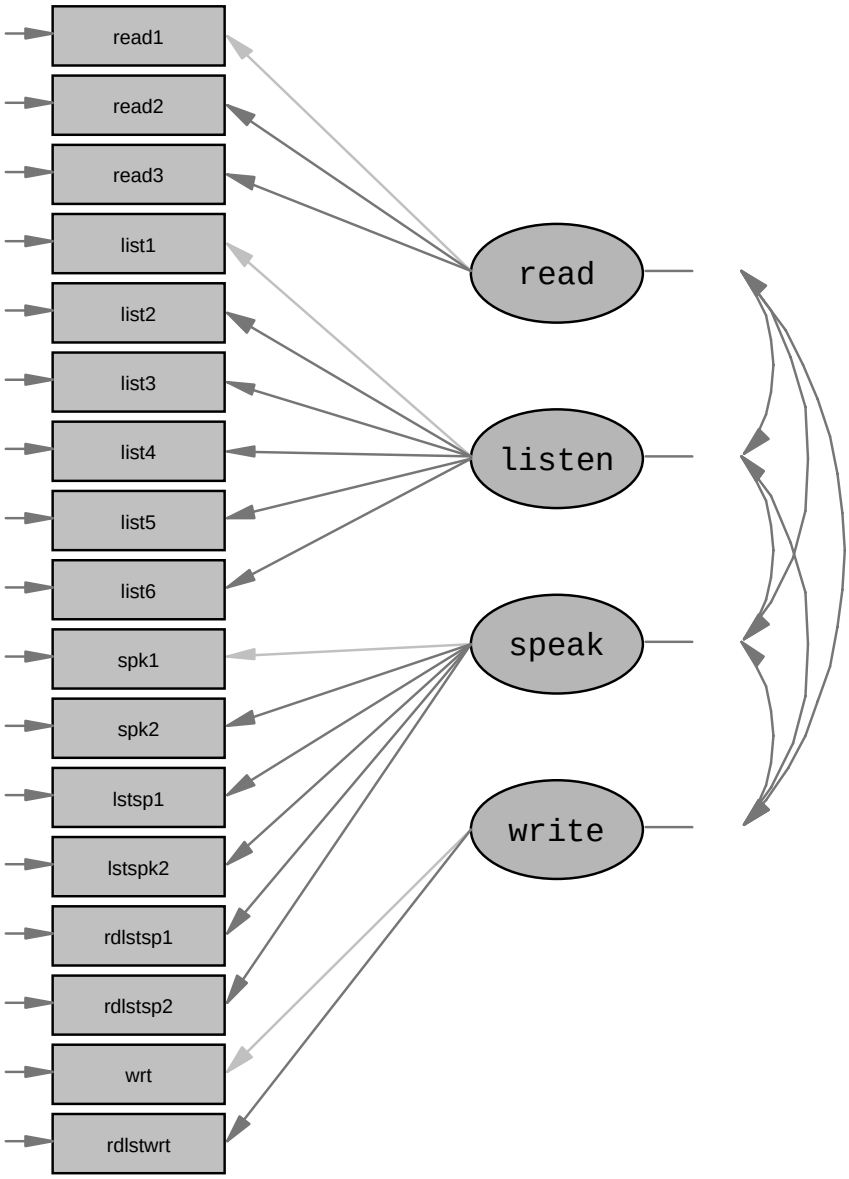

Figure 3. **Model 3: Four factors—Listening, Reading, Speaking, and Writing.**

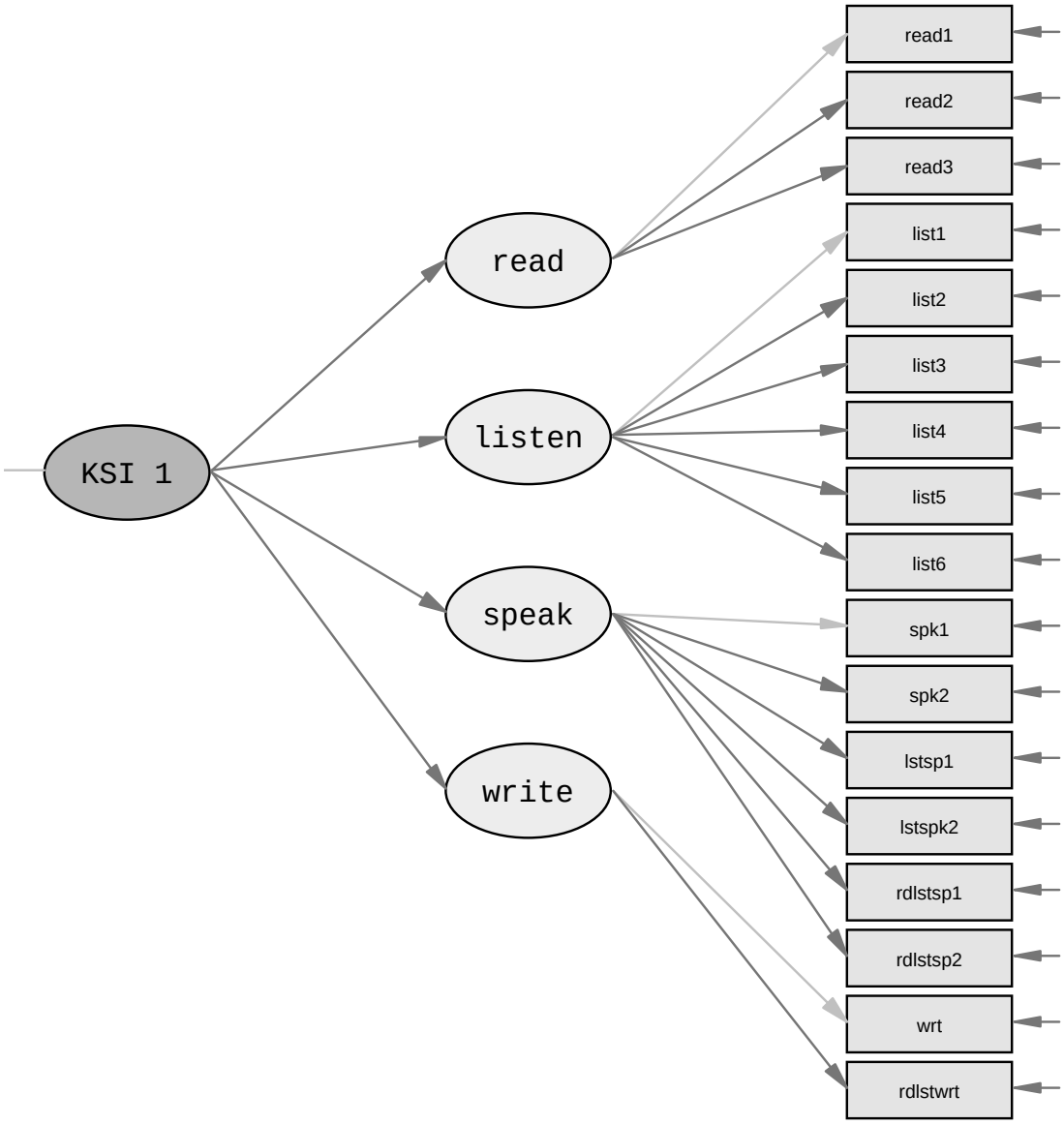

Figure 4. **Model 4: Four first-order factors (Listening, Reading, Speaking, and Writing) and a higher-order factor.**

The nested models for Model 4 were:

1. The number of first-order factors is invariant.
2. The first-order factor loadings are invariant.
3. The first-order factor loadings and error variances are invariant.
4. The first-order factor loadings and error variances, and the higher-order factor loadings are invariant.

The results were evaluated in several ways. Widely used fit indexes were employed (Boomsa, 2000; Hoyle & Panter, 1995; Raykov, Tomer, & Nesselroade, 1991): Satorra-Bentler (S-B) χ^2, S-B χ^2/df, standardized root mean square residual (SRMR), goodness of fit index (GFI), comparative fit index (CFI), nonnormed fit index (NNFI), and root mean square error of approximation (RMSEA). S-B scaled χ^2 difference and S-B scaled χ^2 difference/*df* difference were used in nested comparisons of analyses. The expected cross-validation index (ECVI) was also used in comparisons of overall analyses about the number of factors. In addition, the sizes of factor loadings and factor correlations were examined. The .05 alpha level was used in appraising the S-B χ^2 and S-B χ^2 difference measures, and common rules of thumb were used with the other measures (Hoyle & Panter, 1995; Kline, 1998; Schumaker & Lomax, 1996): 3 or less for S-B χ^2/df and S-B χ^2 difference/ *df* difference; .10 or less for SRMR; .90 or more for GFI, CFI, and NNFI; and .05 or less for RMSEA.

Results and Discussion

Language Family

Four models. The goodness of fit indexes and related information are reported in Table 2 for the four competing models about the number of factors for the language-family samples. (The covariance matrices appear in Appendix A.)

For Model 1 (a single factor), several fit indexes were unsatisfactory for the individual samples and the overall analysis. The GFI was .83 for the Indo-European sample and .78 for the Non-Indo-European sample. And the S-B χ^2/df was 10.66 and the RMSEA was .12 for the overall analysis.

Table 2

Language Family, Tests of Invariance in Number of Factors: Four Models

Sample	*df*	S-B χ^{2a}	S-B χ^2/df	SRMR[b]	GFI[c]	CFI[d]	NNFI[e]	RMSEA[f]	ECVI[g]	Factor correlations exceeding .9
Model 1: One factor—Listening, Reading, Speaking, and Writing										
Indo-European	119			.05	.83					
Non-Indo-European	119			.06	.78					
Overall	238	2,536.20	10.66			.95	.95	.12	2.02	
Model 2: Two factors—Speaking vs. Listening, Reading, and Writing										
Indo-European	118			.03	.94					None
Non-Indo-European	118			.04	.92					None
Overall	236	802.01	3.40			.99	.99	.06	.71	
Model 3: Four factors— Listening, Reading, Speaking, and Writing										
Indo-European	113			.02	.97					.92
Non-Indo-European	113			.02	.96					.91
Overall	226	371.08	1.64			1.00	1.00	.03	.40	--
Model 4: Four first-order factors (Listening, Reading Speaking, and Writing) and a higher-order factor										
Indo-European	115			.03	.96					--
Non-Indo-European	115			.03	.95					--
Overall	230	483.20	2.10			.99	.99	.04	.48	--

[a] Satorra-Bentler χ^2. [b] Standardized root mean square residual. [c] Goodness of fit index. [d] Comparative fit index. [e] Nonnormed fit index. [f] Root mean square error of approximation. [g] Expected cross-validation index.

For Model 2 (two factors, one for Speaking and one for the other sections), two fit indexes were marginally problematic for the overall analysis. The S-B χ^2/df was 3.40, and the RMSEA was .06. There were no high correlations between the factors for either sample.

For Model 3 (four factors), all of the fit indexes for the individual samples and the overall analysis were satisfactory. However, there were marginally problematic correlations between the Reading and Listening factors in both samples: .92 for the Indo-European sample and. 91 for the Non-Indo-European sample.

For Model 4 (four first-order factors and a higher-order factor), all of the fit indexes for the individual samples and the overall analysis were satisfactory.

In short, three of the models seemed acceptable, with satisfactory fits to the data: Model 2 (two factors), Model 3 (four factors), and Model 4 (four first-order factors and a higher-order factor). Detailed comparisons were made of these three models. When Model 2 and Model 3 were compared, the S-B χ^2 difference of 355.66 with 10 *df* (S-B χ^2/df = 35.57) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 3. Furthermore, the RMSEA and ECVI were smaller for Model 3 (.03 vs. .06 for RMSEA and .40 vs. .71 for ECVI). All of these differences indicate better fit for Model 3.

Similarly, when Model 2 and Model 4 were compared, the S-B χ^2 difference of 248.07 with 6 *df* (S-B χ^2/df = 41.35) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 4. And the RMSEA and ECVI were smaller for Model 4 (.04 vs. .06 for RMSEA, and .48 vs. .71 for ECVI). All of these differences indicate a better fit for Model 4.

When Model 3 and Model 4 were compared, the S-B χ^2 difference of 101.41 with 4 *df* (S-B χ^2/df = 25.35) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 3. However, all of the fit indexes were similar for the two models, as were the ECVIs. Given these results, Model 4 was chosen for further analysis because of its parsimony in accounting for the data (including the existence of four distinct first-order factors) with a single, broad higher-order factor.

Model 4 (four first-order factors and a higher-order factor). The goodness of fit indexes and related information are reported in Table 3 for sequential tests about the invariance for the language-family samples in first-order factor loadings and error variances and higher-order factor loadings for Model 4. (Data for the invariance of the number of factors, described already and reported in Table 2, are repeated here for simplicity.)

With regard to the invariance in the number of factors across samples, as already noted, the fit indexes for the individual samples and the overall analysis were satisfactory. This outcome suggests that the number of factors is invariant.

With regard to the invariance in the first-order factor loadings across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis of the number of factors was statistically and practically significant: 64.49 with 13 *df* (S-B $\chi^2/df = 4.96$). However, all of the fit indexes were similar for the two analyses. Taken together, these results suggest that the first-order factor loadings are invariant.

With regard to the invariance in the first-order factor loadings and error variances across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis of first-order factor loadings was statistically and practically significant: 36.74 with 17 *df* (S-B $\chi^2/df =$ 2.16). However, all of the fit indexes were similar for the two analyses. Taken together, these results suggest that the first-order error variances, as well as the factor loadings, are invariant.

With regard to the invariance in the first-order factor loadings and error variances and higher-order factor loadings across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis of first-order factor loadings and error variances was statistically and practically significant: 12.15 with 4 *df* (S-B $\chi^2/df = 3.04$). All of the fit indexes were similar for the two analyses. These results suggest that the higher-order factor loadings, as well as the first-order factor loadings and error variances, are invariant.

The model for invariant first-order factor loadings and error variances and higher-order factor loadings is shown in Figure 5, with common metric, completely standardized factor loadings, and error variances. The first-order factor loadings were substantial (.6 or more), and the higher-order factor loadings were very high (.8 or more). The loadings for the Reading, Listening, and Writing factors on the higher-order factor were noticeably larger (.91 to .97) than the loading for the Speaking factor (.84).

Table 3

Language Family, Complete Tests of Invariance in Factors, Model 4: Four First-Order Factors and a Higher-Order Factor

Sample	*df*	S-B χ^2	S-B χ^2/df	SRMR	GFI	CFI	NNFI	RMSEA
Number of factors invariant								
Indo-European				.03	.96			
Non-Indo-European				.03	.95			
Overall	230	483.20	2.10			.99	.99	.04
First-order factor loadings invariant								
Indo-European				.04	.96			
Non-Indo-European				.04	.95			
Overall	243	542.75	2.23			.99	.99	.04
First-order factor loadings and error variances invariant								
Indo-European				.05	.96			
Non-Indo-European				.05	.94			
Overall	260	579.38	2.23			.99	.99	.04
First-order factor loadings and error variances and higher-order factor loadings invariant								
Indo-European				.06	.95			
Non-Indo-European				.07	.94			
Overall	264	590.99	2.24			.99	.99	.04

Note. See Table 2 footnote for acronyms.

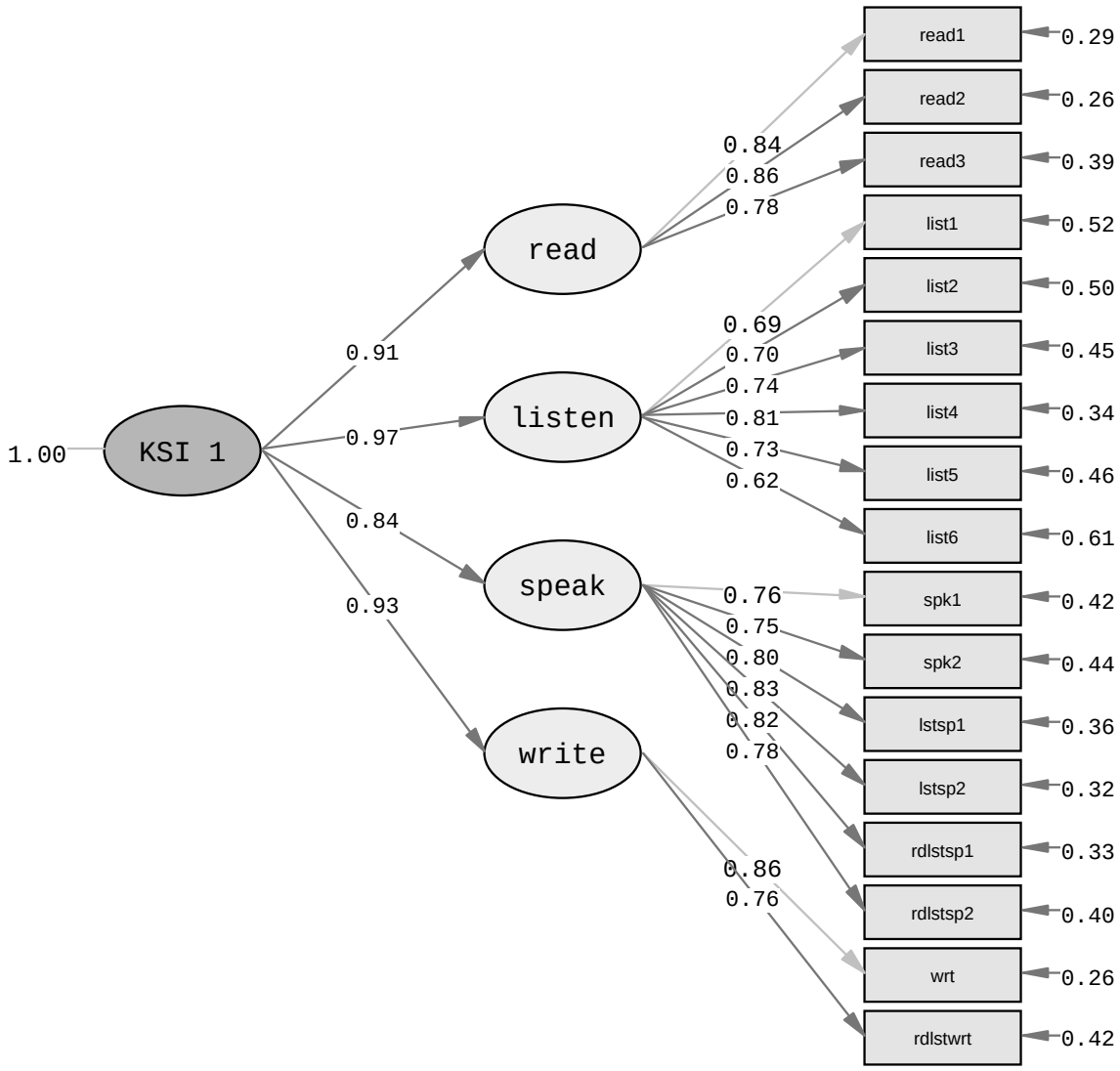

Figure 5. **Language family, Model 4: Four first-order factors (Listening, Reading, Speaking, and Writing) and a higher-order factor. (Common metric, completely standardized factor loadings, and error variances are shown.)**

Outer- and Expanding-Circle Countries

Four models. The goodness of fit indexes and related information are reported in Table 4 for the four competing models about the number of factors for the outer- and expanding-circle countries samples. (The covariance matrices appear in Appendix B.)

For Model 1 (a single factor), several fit indexes were unsatisfactory for the individual samples and the overall analysis. The GFI was .79 for the outer-circle countries sample and .78 for the expanding-circle countries sample. And the S-B χ^2/df was 6.39 and the RMSEA was .13 for the overall analysis.

For Model 2 (two factors, one for Speaking and one for the other sections), two fit indexes were marginally problematic for the individual samples and the overall analysis. The GFI was .89 for the expanding-circle countries sample, and the RMSEA was .07 for the overall analysis. There were no high correlations between the factors for either sample.

For Model 3 (four factors), all of the fit indexes for the individual samples and the overall analysis were satisfactory. However, there were marginally problematic correlations of .93 between the Writing and Listening factors for the outer-circle countries sample, and .94 between the Writing and Speaking factors, and .91 between the Writing and Listening factors for the expanding-circle countries sample.

For Model 4 (four first-order factors and a higher-order factor), all of the fit indexes for the individual samples and the overall analysis were satisfactory.

In short, three of the models seemed acceptable, with satisfactory fits to the data: Model 2 (two factors), Model 3 (four factors), and Model 4 (four first-order factors and a higher-order factor). When Model 2 and Model 3 were compared, the S-B χ^2 difference of 280.98 with 10 *df* (S-B $\chi^2/df = 28.10$) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 3. Furthermore, the RMSEA and ECVI were smaller for Model 3 (.04 vs. .07 for RMSEA, and .71 vs. 1.18 for ECVI). All of the differences indicate a better fit for Model 3.

Similarly, when Model 2 and Model 4 were compared, the S-B χ^2 difference of 198.75 with 6 *df* (S-B $\chi^2/df = 33.12$) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 4. And the RMSEA and ECVI were smaller for Model 4 (.05 vs. .07 for RMSEA, and .83 vs. 1.18 for ECVI). All of these differences indicate a better fit for Model 4.

When Model 3 and Model 4 were compared, the S-B χ^2 difference of 79.37 with 4 *df* (S-B $\chi^2/df = 19.84$) was statistically and practically significant, reflecting a smaller S-B χ^2 for

Table 4

Outer- and Expanding-Circle Countries, Tests of Invariance in Number of Factors: Four Models

Sample	*df*	S-B χ^2	S-B χ^2/df	SRMR	GFI	CFI	NNFI	RMSEA	ECVI	Factor correlations exceeding .9
Model 1: One factor—Listening, Reading, Speaking, and Writing										
Outer circle	119			.06	.79					--
Expanding circle	119			.05	.78					--
Overall	238	1,520.25	6.39			.95	.95	.13	2.41	--
Model 2: Two factors—Speaking vs. Listening, Reading, and Writing										
Outer circle	118			.04	.90					None
Expanding circle	118			.04	.89					None
Overall	236	672.18	2.85			.98	.98	.07	1.18	--
Model 3: Four factors— Listening, Reading, Speaking, and Writing										
Outer circle	113			.03	.94					.93
Expanding circle	113			.02	.95					.91, .94
Overall	226	325.85	1.44			1.00	1.00	.04	.71	--
Model 4: Four first-order factors (Listening, Reading Speaking, and Writing) and a higher-order factor										
Outer circle	115			.04	.93					--
Expanding circle	115			.04	.93					--
Overall	230	418.98	1.82			.99	.99	.05	.83	--

Note. See Table 2 footnote for acronyms.

Model 3. However, all of the fit indexes were similar for the two models, as were the ECVIs. Based on these results and parsimony, Model 4 was chosen for further analysis.

Model 4 (four first-order factors and a higher-order factor). The goodness of fit indexes and related information are reported in Table 5 for sequential tests about the invariance for the outer- and expanding-circle countries samples in the parameters for Model 4.

With regard to the invariance in the number of factors across samples, the fit indexes for the individual samples and the overall analysis were satisfactory. This outcome suggests that the number of factors is invariant.

With regard to the invariance in the first-order factor loadings across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was not statistically significant: 22.23 with 13 *df* (S-B χ^2/df = 1.71). All of the fit indexes were similar for the two analyses. These results suggest that the first-order factor loadings are invariant.

With regard to the invariance in the first-order factor loadings and error variances across samples, all of the fit indexes for the individual samples and the overall analyses were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was not statistically significant: 23.83 with 17 *df* (S-B χ^2/df = 1.40). All of the fit indexes were similar for the two analyses. These results suggest that the first-order error variances, as well as the factor loadings, are invariant.

With regard to the invariance in the first-order factor loadings and error variances, and the higher-order factor loadings across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was not statistically significant: 8.30 with 4 *df* (S-B χ^2/df = 2.08). All of the fit indexes were similar for the two analyses. These results suggest that the higher-order factor loadings, as well as the first-order factor loadings and error variances, are invariant.

The model for invariant first-order factor loadings and error variances and higher-order factor loadings is shown in Figure 6, with common metric, completely standardized factor loadings and error variances. The first-order factor loadings were substantial, and the higher-order factor loadings were very high. The loadings for the Listening and Writing factors on the higher-order factor were noticeably larger (.96 and .95, respectively) than the loadings for the Reading and Speaking factors (.89 and .87, respectively).

Table 5

Outer- and Expanding-Circle Countries, Complete Tests of Invariance in Factors, Model 4: Four First-Order Factors and a Higher-Order Factor

Sample	*df*	S-B χ^2	S-B χ^2/df	SRMR	GFI	CFI	NNFI	RMSEA
Number of factors invariant								
Outer circle				.04	.93			
Expanding circle				.04	.93			
Overall	230	418.98	1.82			.99	.99	05
First-order factor loadings invariant								
Outer circle				.05	.93			
Expanding circle				.04	.93			
Overall	243	441.44	1.82			.99	.99	.05
First-order factor loadings and error variances invariant								
Outer circle				.05	.92			
Expanding circle				.04	.93			
Overall	260	463.94	1.78			.99	.99	.05
First-order factor loadings and error variances and higher-order factor loadings invariant								
Outer circle				.07	.92			
Expanding circle				.06	.93			
Overall	264	476.62	1.81			.99	.99	.05

Note. See Table 2 footnote for acronyms.

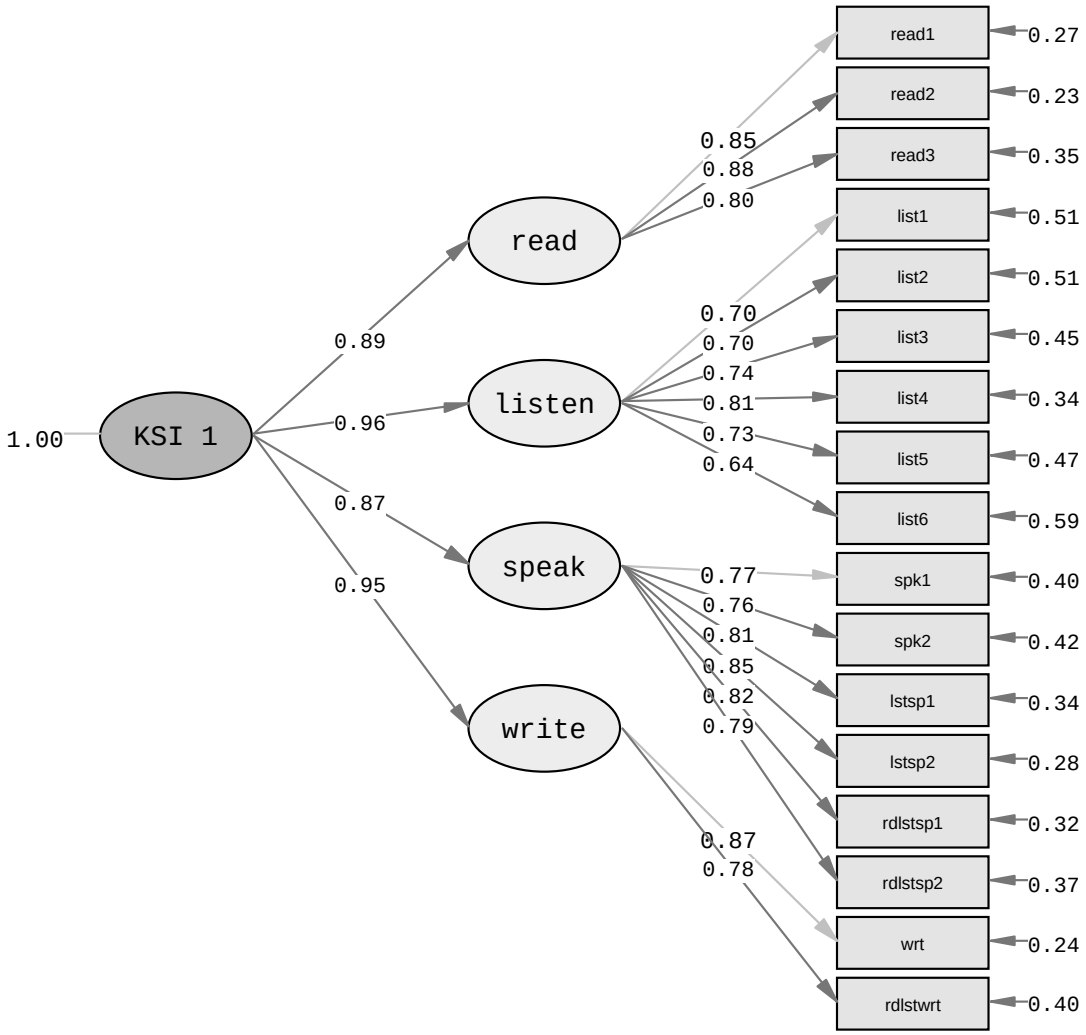

Figure 6. **Outer- and expanding-circle countries, Model 4: Four first-order factors (Listening, Reading, Speaking, and Writing) and a higher-order factor. (Common metric, completely standardized factor loadings, and error variances are shown.)**

English-Language Study

Four models. The goodness of fit indexes and related information are reported in Table 6 for the four competing models about the number of factors for the English-language study samples. (The covariance matrices appear in Appendix C.)

For Model 1 (a single factor), several fit indexes were unsatisfactory for the individual samples and the overall analysis. The GFI was .78 for the 6 years or less sample, .79 for the 7 to 10 years sample, and .83 for the 11 years or more sample. And the S-B χ^2/df was 7.83 and the RMSEA was .12 for the overall analysis.

For Model 2 (two factors, one for Speaking and one for the other sections), one fit index was marginally problematic for the overall analysis; the RMSEA was .06. There were no high correlations between the factors in the individual samples.

For Model 3 (four factors), all of the fit indexes for the individual samples and the overall analysis were satisfactory. However, there were marginally problematic correlations between the Listening and Reading factors for both the 7 to 10 years sample (.92) and the 11 years or more sample (.94).

For Model 4 (four first-order factors and a higher-order factor), all of the fit indexes for the individual samples and the overall analysis were satisfactory.

In short, three of the models seemed acceptable: Model 2 (two factors), Model 3 (four factors), and Model 4 (four first-order factors and a higher-order factor). When Model 2 and Model 3 were compared, the S-B χ^2 difference of 360.59 with 15 *df* (S-B χ^2/df = 24.04) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 3. Furthermore, the RMSEA and ECVI were smaller for Model 3 (.03 vs. .06 for RMSEA and .53 vs. .84 for ECVI). All of these differences indicate a better fit for Model 3.

Similarly, when Model 2 and Model 4 were compared, the S-B χ^2 difference of 240.89 with 9 *df* (S-B χ^2/df = 26.77) was statistically and practically significant, reflecting a smaller S-Bχ^2 for Model 4. And the RMSEA and ECVI were smaller for Model 4 (.04 vs. .06 for the former and .61 vs. .84 for the latter). All of these differences indicate a better fit for Model 4.

When Model 3 and Model 4 were compared, the S-B χ^2 difference of 112.42 with 6 *df* (S-B χ^2/df = 18.74) was statistically and practically significant, reflecting a smaller S-B χ^2 for Model 3. However, all of the fit indexes were similar for the two models, as were the ECVIs. Based on these results and parsimony, Model 4 was chosen for further analysis.

Table 6

English-Language Study, Tests of Invariance in Number of Factors: Four Models

Sample	*df*	S-B χ^2	S-B χ^2/df	SRMR	GFI	CFI	NNFI	RMSEA	ECVI	Factor correlations exceeding .9
Model 1: One factor—Listening, Reading, Speaking, and Writing										
6 years or less	119			.06	.78					--
7 to 10 years	119			.06	.79					--
11 years or more	119			.05	.83					--
Overall	357	2797.06	7.83			.95	.94	.12	2.15	--
Model 2: Two factors—Speaking vs. Listening, Reading, and Writing										
6 years or less	118			.04	.92					None
7 to 10 years	118			.04	.92					None
11 years or more	118			.04	.92					None
Overall	354	964.67	2.73			.99	.99	.06	.84	--
Model 3: Four factors— Listening, Reading, Speaking, and Writing										
6 years or less	113			.03	.96					None
7 to 10 years	113			.02	.96					.92
11 years or more	113			.03	.95					.94
Overall	339	505.68	1.49			1.00	1.00	.03	.53	--
Model 4: Four first-order factors (Listening, Reading Speaking, and Writing) and a higher-order factor										
6 years or less	115			.03	.95					--
7 to 10 years	115			.03	.95					--
11 years or more	115			.03	.94					--
Overall	345	629.28	1.82			.99	.99	.04	.61	--

Note. See Table 2 footnote for acronyms.

Model 4 (four first-order factors and a higher-order factor). The goodness of fit indexes and related information are reported in Table 7 for sequential tests about the invariance for the English-language study samples in the parameters for Model 4.

With regard to the invariance in the number of factors across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. This outcome suggests that the number of factors is invariant.

With regard to the invariance in the first-order factor loadings on the first-order factors across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was statistically and practically significant: 141.73 with 26 *df* (S-B $\chi^2/df = 5.45$). However, all of the fit indexes were similar for the two analyses. Taken together, these results suggest that the factor loadings are invariant.

With regard to the invariance in the first-order factor loadings and error variances across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was statistically but not practically significant: 81.27 with 34 *df* (S-B $\chi^2/df = 2.39$). All of the fit indexes were similar for the two analyses. These results suggest that the first-order error variances, as well as the factor loadings, are invariant.

With regard to the invariance in the first-order factor loadings and error variances and higher-order factor loadings across samples, all of the fit indexes for the individual samples and the overall analysis were satisfactory. The S-B χ^2 difference between this overall analysis and the preceding overall analysis was statistically and practically significant: 27.79 with 8 *df* (*S-B* $\chi^2/df = 3.47$). However, all of the fit indexes were similar for the two analyses. Taken together, these results suggest that the higher-order factor loadings, as well as the first-order factor loadings and error variances, are invariant.

The model for invariant first-order loadings and error variances and higher-order factor loadings is shown in Figure 7, with common metric, completely standardized factor loadings, and error variances. The first-order factor loadings were substantial, and the higher-order factor loadings were very high. The loadings for the Reading, Listening, and Writing factors on the higher-order factor were noticeably larger (.90 to .97) than the loading for the Speaking factor (.83).

Table 7

English-Language Study, Complete Tests of Invariance in Factors, Model 4: Four First-Order Factors and a Higher-Order Factor

Sample	*df*	S-B χ^2	S-B χ^2/df	SRMR	GFI	CFI	NNFI	RMSEA
Number of factors invariant								
6 years or less				.03	.95			
7 to 10 years				.03	.95			
11 years or more				.03	.94			
Overall	345	629.28	1.82			.99	.99	.04
First-order factor loadings invariant								
6 years or less				.06	.94			
7 to 10 years				.04	.95			
11 years or more				.07	.93			
Overall	371	756.79	2.04			.99	.99	.05
First-order factor loadings and error variances invariant								
6 years or less				.07	.93			
7 to 10 years				.05	.94			
11 years or more				.07	.92			
Overall	405	837.96	2.07			.99	.99	.05
First-order factor loadings and error variances, and higher-order factor loadings invariant								
6 years or less				.07	.93			
7 to 10 years				.05	.94			
11 years or more				.08	.92			
Overall	413	863.66	2.09			.99	.99	.05

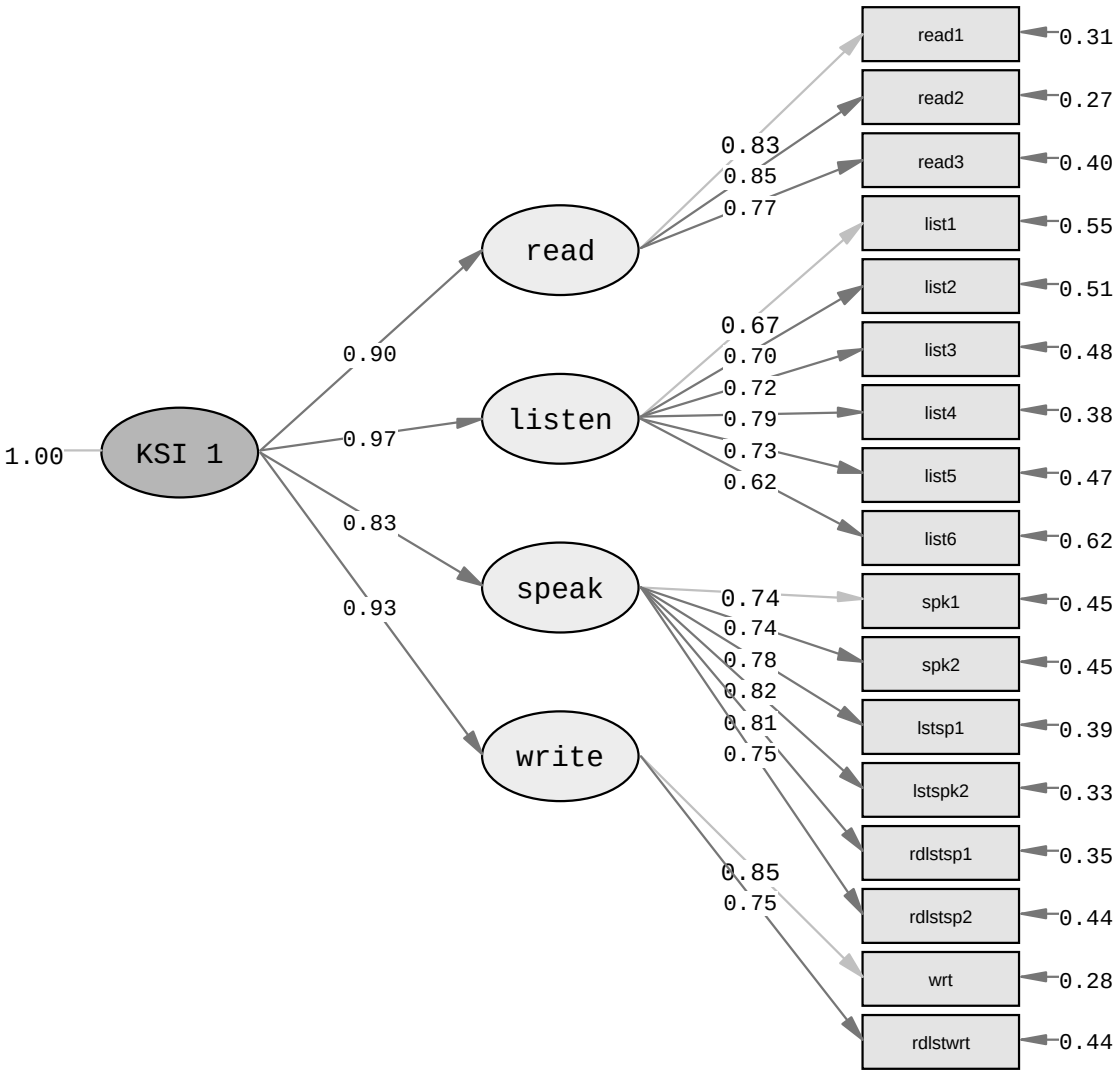

Figure 7. **English-language study, Model 4: Four first-order factors (Listening, Reading, Speaking, and Writing) and a higher-order factor. (Common metric, completely standardized factor loadings, and error variances are shown.)**

Conclusion

The consistency of the findings across the three populations of test takers is remarkable. In each population and in each subgroup within these populations, the same factor-analytic model was identified on the basis of model fit and parsimony, with one higher-order factor subsuming four correlated first-order factors that correspond to the sections of the TOEFL iBT assessment. (The parsimony of a higher-order solution is reinforced by the high intercorrelations, in the .6 to .9 range, between the first-order factors.)[4]

This higher-order factor model, congruent with the consensus that language ability is multicomponential (e.g., Sasaki, 1999), is identical to the one identified by Sawaki et al. (2008), using the same data set, despite substantial differences in the groups of test takers (Sawaki et al. used the total sample; this study used subsamples likely to diverge) and analytic methods (Sawaki et al. analyzed individual items; this study analyzed parcels of items). However, this model differs from the one identified by Stricker et al. (2005) in a study of a TOEFL iBT prototype: two correlated factors, one for the Speaking section and one for the Listening, Reading, and Writing sections. The reason for this divergence is unclear. One conjecture is that three language groups (Arabic, Chinese, and Spanish) in the Stricker et al. investigation may be more homogeneous in educational background and experience than the subgroups in the present study. It seems unlikely that the relatively minor differences between the TOEFL iBT prototype in their study and the final version of the test in the present study is responsible.

The present finding has similarities and differences with the results of previous studies that factor analyzed other ESL tests of the same four language skills (Bachman, Davidson, Ryan, & Choi, 1995; Carroll, 1983; Kunnan, 1995; Shin, 2005). On the one hand, all of the studies found three factors: Speaking, Listening, and a fusion of Reading and Writing. On the other hand, Carroll, Bachman et al., and Shin found a higher-order factor defined by these first-order factors. However, the generalizability of these studies is circumscribed, for all but one (Carroll) are based on portions of the same data set.

The invariance across subgroups in this study accords with the invariance across language groups in the Stricker et al. (2005) study of the TOEFL iBT prototype but is contrary to the expectation that the factor structure might vary for the subgroups, depending on their language family and exposure to English. However, the potential influence of these variables on the functioning of the TOEFL iBT assessment cannot necessarily be ruled out on the basis of

these findings. The research participants in this study, in common with actual TOEFL test takers, are necessarily atypical of ESL speakers in their parent populations, for they clearly have confidence to take a high-level test of their proficiency in the English language and may also have greater English-language skills than their fellow ESL speakers.

The invariance in this study is broadly consistent with the findings of two previous studies of other ESL tests, cited earlier. Both Kunnan (1995) and Shin (2005) observed some invariance in the three language factors that they identified, Kunnan investigating invariance across Indo-European and Non-Indo-European language family samples, and Shin investigating samples that varied in their level of English-language proficiency as defined by other ESL tests.

One implication of these results for the TOEFL iBT assessment is the same as the one already noted by Sawaki et al. (2008). The higher-order structure that was identified supports the current scoring of the test: separate scores for each section, and a total score. (The generally lower loadings of the Speaking factor on the higher-order factor, also observed by Sawaki et al., is somewhat problematic for the current total score.) Another implication is that the test is functioning in the same way across diverse subgroups of test takers.

It should be borne in mind that these findings are based on data for research participants. Although they were recruited to reflect the TOEFL test-taking population, it is uncertain how representative these participants are in their educational background and their test-taking motivation. A follow-up study with actual test takers is needed. Such a study could also assess invariance across specific language groups, which was precluded in the present study because of inadequate sample sizes. Another caveat is in order: The findings about Kachru's (1984, 1985) outer- and expanding-circle countries should not be overinterpreted, for the classification of expanding-circle countries that was used may be obsolete, given the increased use of ESL across the world in the two decades since this scheme was formulated (e.g., Crystal, 1997; Graddol, 1997).

References

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*. Cambridge, England: Cambridge University Press.

Boomsa, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*, 461–483.

Bruthiaux, P. (2003). Squaring the circles: Issues in modeling English worldwide. *International Journal of Applied Linguistics*, *13*, 159–178.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr., (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Crystal, D. (1997). *English as a global language*. Cambridge, England: Cambridge University Press.

ETS. (2002). *LanguEdge courseware score interpretation guide.* Princeton, NJ: Author.

ETS. (2004). *The next generation TOEFL test: Focus on communication*. Retrieved April 26, 2004, from http://www.ets.org/toefl/nextgen

Gorsuch, R. (1974). *Factor analysis*. Philadelphia: Saunders.

Graddol, D., (1997). *The future of English*? *A guide to forecasting the popularity of the English language in the 21st century*. London: British Council.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling—Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Joreskog, K. G., & Sorbom, D. (1996a). LISREL 8: User's reference guide [Computer software manual]. Chicago: Scientific Software.

Joreskog, K. G., & Sorbom, D. (1996b). PRELIS 2: User's reference guide [Computer software manual]. Chicago: Scientific Software.

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

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world:*

Teaching and learning the language and literatures (pp. 11–30). Cambridge, England: Cambridge University Press.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

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

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in *Psychology and Aging*: Some proposed guidelines. *Psychology and Aging*, *6*, 499–503.

Sasaki, M. (1999). *Second language proficiency, foreign language aptitude, and intelligence—Quantitative and qualitative analyses*. New York: Lang.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Explorations in a field trial sample* (TOEFL iBT Research Rep. No. 4; ETS Research Rep. No. 08-09). Princeton, NJ: ETS.

Schumaker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

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., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series Rep. No. MS-32: ETS Research Rep. No. 05-12). Princeton, NJ: ETS.

Thumboo, E. (Ed.) (2001). *The three circles of English: Language specialists talk about the English language*. Singapore: Unipress, National University of Singapore.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chappelle, M. K. Enright, & J. M. Jamison (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 259–318). New York: Routledge.

Xi, X., Midouhas, E., & Steinberg, J. (2006). *Language learning backgrounds represented by TOEFL iBT test takers and their impact on TOEFL iBT performance*. Manuscript in preparation.

Yano, Y. (2001). World Englishes in 2000 and beyond. *World Englishes*, *20*, 119–131.

Notes

[1] The Indo-European countries were Afghanistan, Argentina, Bangladesh, Bulgaria, Brazil, Colombia, Croatia, France, Germany, Greece, India, Italy, Jamaica, Mexico, Norway, Pakistan, Poland, Romania, Russia, and Yugoslavia. The non-Indo-European countries were Algeria, China, Egypt, Ethiopia, Hong Kong, Indonesia, Jordan, Japan, Korea, Kuwait, Lebanon, Malaysia, Morocco, the Philippines, Saudi Arabia, Senegal, Taiwan, Thailand, Turkey, the United Arab Republic, Vietnam, and the West Bank.

[2] Kachru's (1984, 1985) original classification of outer- and expanding-circle countries was used, in view of the debate (e.g., Bruthiaux, 2003; Yano, 2001) about subsequent extensions of this scheme (Crystal, 1997; Graddol, 1997). Outer-circle countries were India, Malaysia, Pakistan, and the Philippines. Expanding-circle countries were China, Egypt, Indonesia, Japan, Korea, Russia, Saudi Arabia, and Taiwan. (Inner -circle countries—Australia, Canada, Great Britain, the United States, and New Zealand—were excluded because the TOEFL assessment is not ordinarily taken by test takers from these countries.)

[3] Years of English-language study were grouped in part on the basis of the finding that TOEFL iBT scores were similar for the first five or six years of instruction (Wang et al., 2008).

[4] In the initial analyses of the number of invariant factors for Model 3 (four factors), the factor intercorrelations ranged from .74 to .92 for the Indo-European sample and from .70 to .91 for the Non-Indo-European sample; from .75 to .93 for the outer-circle countries sample, from .69 to .94 for the expanding-circle countries sample; and from .66 to .89 for the 6 years or less of English-language study sample, from .68 to .92 for the 7 to 10 years sample, and from .76 to .94 for the 11 years or more sample.

List of Appendixes

Appendix A

Covariance Matrices for Language Family Samples

Table A1

Covariance Matrix for Indo-European Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.17																
2. Reading—Prompt 2	6.85	8.71															
3. Reading—Prompt 3	7.26	6.99	12.21														
4. Listening—Prompt 1	2.45	2.24	2.62	2.18													
5. Listening—Prompt 2	1.78	1.79	1.87	.85	1.37												
6. Listening—Prompt 3	3.17	3.04	3.44	1.31	.93	2.99											
7. Listening—Prompt 4	3.62	3.61	3.75	1.52	1.19	1.96	3.44										
8. Listening—Prompt 5	2.36	2.27	2.32	.84	.77	1.18	1.39	1.77									
9. Listening—Prompt 6	2.23	2.05	2.29	.89	.67	1.20	1.40	.91	1.92								
10. Speaking—Task 1	1.34	1.32	1.34	.54	.43	.73	.85	.47	.44	.95							
11. Speaking—Task 2	1.19	1.16	1.29	.50	.43	.70	.79	.46	.41	.52	.94						
12. Listening/Speaking 1	11.84	1.57	1.69	.67	.49	.91	1.03	.60	.52	.58	.57	1.17					
13. Listening/Speaking 2	21.83	1.85	1.93	.77	.55	1.00	1.14	.72	.65	.65	.62	.79	1.37				
14. Reading/Listening/Speaking 1	1.61	1.48	1.65	.67	.55	.81	.98	.60	.48	.58	57	.69	.76	1.11			
15. Reading/Listening/Speaking 2	1.72	1.62	1.73	.69	.54	.85	1.05	.61	.61	.55	.52	.66	.77	.67	1.16		
16. Writing	2.66	2.53	2.98	1.19	.86	1.40	1.65	.93	.91	.67	.63	.85	.96	.82	.86	2.03	
17. Reading/Listening/Writing	2.04	1.80	2.23	.78	.62	1.07	1.17	.71	.66	.58	.57	.60	.75	.65	.67	1.12	1.45

Table A2

Covariance Matrix for Non-Indo-European Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.54																
2. Reading—Prompt 2	7.59	9.86															
3. Reading—Prompt 3	7.21	7.48	12.46														
4. Listening—Prompt 1	2.58	2.60	2.47	2.16													
5. Listening—Prompt 2	2.44	2.32	2.32	1.08	1.80												
6. Listening—Prompt 3	2.90	3.05	3.16	1.26	1.17	2.80											
7. Listening—Prompt 4	3.95	3.67	3.72	1.60	1.47	1.84	3.50										
8. Listening—Prompt 5	3.04	3.12	2.95	1.35	1.26	1.53	1.86	2.62									
9. Listening—Prompt 6	2.32	2.12	2.24	.71	.76	1.15	1.24	1.07	1.89								
10. Speaking—Task 1	1.66	1.54	1.77	.74	.71	.82	.99	.82	.54	1.12							
11. Speaking—Task 2	1.50	1.44	1.70	.66	.64	.84	.92	.78	.55	.70	.95						
12. Listening/Speaking 1	1.80	1.67	2.08	.82	.74	.95	1.17	.96	.69	.76	.70	1.22					
13. Listening/Speaking 2	1.99	1.88	2.20	.88	.87	1.10	1.30	1.09	.72	.89	.82	.97	1.56				
14. Reading/Listening/Speaking 1	1.84	1.83	2.02	.88	.88	1.05	1.18	1.04	.65	.81	.79	.89	1.04	1.40			
15. Reading/Listening/Speaking 2	1.64	1.65	1.85	.75	.77	.89	1.07	.90	.61	.75	.65	.80	.97	.95	1.26		
16. Writing	2.88	2.69	3.03	1.09	1.06	1.32	1.60	1.28	.92	.84	.82	.98	1.08	1.00	.91	1.94	
17. Reading/Listening/Writing	2.52	2.34	2.72	.89	.99	1.10	1.36	1.12	.73	.77	.72	.89	1.01	.89	.85	1.26	1.84

Appendix B

Covariance Matrices for Outer- and Expanding-Circle Countries Samples

Table B1

Covariance Matrix for Outer-Circle Countries Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.57																
2. Reading—Prompt 2	7.52	9.32															
3. Reading—Prompt 3	7.42	7.68	12.02														
4. Listening—Prompt 1	2.45	2.14	2.57	1.83													
5. Listening—Prompt 2	1.92	2.03	2.21	.74	1.51												
6. Listening—Prompt 3	2.87	2.86	3.46	1.16	.88	2.64											
7. Listening—Prompt 4	3.74	3.53	3.80	1.39	1.21	1.80	3.33										
8. Listening—Prompt 5	2.60	2.58	2.68	.81	.91	1.03	1.43	2.01									
9. Listening—Prompt 6	2.42	2.04	2.53	.78	.73	1.06	1.36	.96	1.85								
10. Speaking—Task 1	1.31	1.46	1.58	.45	.45	.67	.86	.51	.46	1.11							
11. Speaking—Task 2	1.31	1.37	1.66	.48	.47	.66	.88	.56	.48	.67	1.04						
12. Listening/Speaking 1	1.82	1.79	2.12	.65	.62	.92	1.07	.63	.63	.70	.66	1.29					
13. Listening/Speaking 2	1.89	2.08	2.29	.68	.59	.92	1.15	.77	.70	.76	.73	.86	1.37				
14. Reading/Listening/Speaking 1	1.58	1.74	2.11	.61	.61	.75	.97	.68	.50	.66	.65	.73	.84	1.20			
15. Reading/Listening/Speaking 2	1.76	1.76	2.05	.59	.52	.74	.98	.62	.65	.63	.58	.77	.82	.75	1.23		
16. Writing	3.01	2.83	3.59	1.28	1.01	1.48	1.77	1.14	1.13	.84	.86	1.04	1.11	1.03	1.00	2.43	
17. Reading/Listening/Writing	2.25	2.09	2.73	.86	.72	1.20	1.37	.91	.87	.79	.75	.76	.86	.84	.81	1.53	1.86

Table B2

Covariance Matrix for Expanding-Circle Countries Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.70																
2. Reading—Prompt 2	8.03	10.15															
3. Reading—Prompt 3	7.49	7.61	12.11														
4. Listening—Prompt 1	2.55	2.57	2.51	2.15													
5. Listening—Prompt 2	2.31	2.33	2.28	1.11	1.89												
6. Listening—Prompt 3	2.88	3.21	3.19	1.34	1.21	3.02											
7. Listening—Prompt 4	4.02	3.89	3.81	1.69	1.51	2.07	3.56										
8. Listening—Prompt 5	2.89	3.09	2.85	1.34	1.25	1.53	1.82	2.51									
9. Listening—Prompt 6	2.37	2.22	2.09	.75	.75	1.21	1.31	.98	1.83								
10. Speaking—Task 1	1.72	1.57	1.72	.72	.70	.92	1.08	.79	.63	1.10							
11. Speaking—Task 2	1.47	1.50	1.66	.70	.71	.94	1.00	.76.	59	.69	.94						
12. Listening/Speaking 1	1.81	1.68	1.86	.86	.76	1.08	1.25	.97	.69	.78	.70	1.23					
13. Listening/Speaking 2	2.14	2.09	2.09	.96	.97	1.32	1.44	1.14	.79	.92	.81	1.02	1.64				
14. Reading/Listening/Speaking 1	1.98	1.87	1.88	.98	.99	1.25.	1.32	1.11	.74	.82	.80	.95	1.11	1.43			
15. Reading/Listening/Speaking 2	1.78	1.80	1.79	.83	.85	1.07	1.21	.97	.62	.76	.65	.83	1.00	.97	1.26		
16. Writing	3.00	2.76	2.78	1.08	1.10	1.38	1.69	1.26	.93	.88	.81	1.05	1.18	1.08	1.00	1.90	
17. Reading/Listening/Writing	2.54	2.42	2.56	.86	.99	1.15	1.31	1.10	.76	.81	.73	.88	1.09	.91	.89	1.16	1.72

Appendix C

Covariance Matrices for English-Language Study Samples

Table C1

Covariance Matrix for 6 Years or Less Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.20																
2. Reading—Prompt 2	7.12	9.60															
3. Reading—Prompt 3	6.27	6.46	11.02														
4. Listening—Prompt 1	2.21	2.13	2.04	2.06													
5. Listening—Prompt 2	2.10	2.03	1.84	1.01	1.79												
6. Listening—Prompt 3	2.56	2.58	2.79	1.13	1.05	2.69											
7. Listening—Prompt 4	3.37	3.40	3.18	1.38	1.33	1.69	3.27										
8. Listening—Prompt 5	3.12	3.13	2.71	1.23	1.27	1.43	1.83	2.77									
9. Listening—Prompt 6	2.09	1.89	1.85	.71	.69	1.01	1.14	1.22	1.85								
10. Speaking—Task 1	1.38	1.18	1.16	.52	.57	.67	.76	.66	.42	.92							
11. Speaking—Task 2	1.19	1.13	1.19	.47	.57	.63	.69	.69	.41	.54	.89						
12. Listening/Speaking 1	1.51	1.40	1.48	.56	.58	.78	.89	.86	.57	.58	.58	1.09					
13. Listening/Speaking 2	1.78	1.63	1.61	.68	.79	.89	1.05	1.04	.58	.73	.65	.82	1.45				
14. Reading/Listening/Speaking 1	1.77	1.68	1.73	.76	.86	.91	1.09	1.00	.59	.70	.69	.81	.96	1.38			
15. Reading/Listening/Speaking 2	1.49	1.33	1.35	.57	.73	.71	.90	.86	.51	.57	.52	.67	.84	.84	1.21		
16. Writing	2.14	2.06	1.94	.81	.85	1.00	1.28	1.08	.71	.55	.50	.67	.81	.79	.67	1.37	
17. Reading/Listening/Writing	2.11	1.98	2.11	.67	.88	.90	1.18	1.03	.59	.63	.60	.72	.87	.87	.69	.89	1.58

Table C2

Covariance Matrix for 7 to 10 Years Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	10.04																
2. Reading—Prompt 2	6.79	8.49															
3. Reading—Prompt 3	7.10	6.65	11.28														
4. Listening—Prompt 1	2.36	2.41	2.52	2.15													
5. Listening—Prompt 2	2.09	1.94	1.97	.91	1.60												
6. Listening—Prompt 3	2.96	3.09	2.99	1.22	.99	2.96											
7. Listening—Prompt 4	3.59	3.38	3.25	1.40	1.29	1.82	3.31										
8. Listening—Prompt 5	2.40	2.42	2.25	1.01	.97	1.22	1.35	1.95									
9. Listening—Prompt 6	2.14	2.05	2.16	.74	.71	1.13	1.28	.86	1.79								
10. Speaking—Task 1	1.18	1.21	1.17	.59	.52	.70	.79	.58	.45	.93							
11. Speaking—Task 2	1.34	1.28	1.28	.66	.52	.76	.83	.55	.47	.54	.92						
12. Listening/Speaking 1	1.84	1.60	1.60	.73	.63	.90	1.16	.76	.57	.63	.61	1.19					
13. Listening/Speaking 2	1.59	1.58	1.47	.76	.65	.93	1.07	.72	.68	.66	.69	.77	1.30				
14. Reading/Listening/Speaking 1	1.43	1.36	1.27	.67	.62	.82	.90	.71	.51	.61	.63	.69	.74	1.08			
15. Reading/Listening/Speaking 2	1.39	1.45	1.43	.67	.66	.84	.93	.66	.57	.56	.56	.63	.75	.65	1.00		
16. Writing	2.81	2.52	2.86	1.10	.96	1.31	1.57	1.03	.88	.71	.79	.92	.91	.79	.83	2.02	
17. Reading/Listening/Writing	2.17	1.92	2.11	.82	.77	1.00	1.11	.83	.66	.54	.60	.65	.67	.56	.65	1.10	1.43

Table C3

Covariance Matrix for 11 Years or More Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Reading—Prompt 1	9.36																
2. Reading—Prompt 2	6.05	7.83															
3. Reading—Prompt 3	6.96	6.97	12.77														
4. Listening—Prompt 1	2.13	1.89	2.57	1.86													
5. Listening—Prompt 2	1.76	1.73	2.07	.68	1.25												
6. Listening—Prompt 3	2.78	2.65	3.56	1.12	.85	2.67											
7. Listening—Prompt 4	3.32	3.06	3.58	1.39	1.04	1.66	3.14										
8. Listening—Prompt 5	1.98	1.83	2.24	.67	.64	1.04	1.16	1.51									
9. Listening—Prompt 6	2.16	1.83	2.52	.76	.70	1.09	1.23	.74	1.94								
10. Speaking—Task 1	1.40	1.35	1.76	.55	.51	.72	.95	.46	.48	1.09							
11. Speaking—Task 2	1.20	1.08	1.61	.46	.43	.68	.84	.48	.44	.62	.90						
12. Listening/Speaking 1	1.75	1.45	2.01	.69	.60	.91	.98	.55	.59	.62	.58	1.18					
13. Listening/Speaking 2	1.75	1.75	2.26	.71	.64	1.00	1.14	.69	.70	.75	.67	.80	1.35				
14. Reading/Listening/Speaking 1	1.39	1.35	1.82	.61	.50	.77	.87	.51	.44	.56	.56	.67	.77	1.03			
15. Reading/Listening/Speaking 2	1.38	1.37	1.72	.56	.45	.73	.84	.46	.54	.56	.48	.64	.72	.65	1.08		
16. Writing	2.70	2.51	3.68	1.12	.88	1.40	1.54	.87	1.03	.79	.78	.94	1.03	.85	.86	2.23	
17. Reading/Listening/Writing	1.95	1.70	2.56	.69	.62	1.09	1.13	.63	.68	.67	.62	.69	.84	.64	.63	1.30	1.62

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands