ISSN 1556-9012

ETS® TOEFL®

Monograph Series

MS- 32
September 2005

Factor Structure of the LanguEdge™ Test Across Language Groups

Lawrence J. Stricker
Donald A. Rock
Yong-Won Lee

Factor Structure of the LanguEdge™ Test Across Language Groups

Lawrence J. Stricker, Donald A. Rock, and Yong-Won Lee
ETS, Princeton, NJ

RR-05-12

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 © 2005 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.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, Graduate Record Examinations, GRE, TOEFL, the TOEFL logo, and TSE are registered trademarks of Educational Testing Service. LanguEdge, The Test of English as a Foreign Language, and the Test of Spoken English are trademarks of Educational Testing Service.

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

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language™ (TOEFL®) test development efforts. As part of the foundation for the development of the next generation TOEFL test, papers and research reports were commissioned from experts within the fields of measurement, language teaching, and testing through the TOEFL 2000 project. The resulting critical reviews, expert opinions, and research results have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project was a broad effort under which language testing at Educational Testing Service® (ETS®) would evolve into the 21st century. As a first step, the TOEFL program revised the Test of Spoken English™ (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, took advantage of new forms of assessment and improved services made possible by computer-based testing, while also moving the program toward its longer-range goals, which included:

- the development of a conceptual framework that takes into account models of communicative competence
- a research program that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model. The culmination of the TOEFL 2000 project is the next generation TOEFL test that will be released in September 2005.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program
Educational Testing Service

Abstract

This study assessed the factor structure of the LanguEdge™ test and the invariance of its factors across language groups. Confirmatory factor analyses of individual tasks and subsets of items in the four sections of the test, Listening, Reading, Speaking, and Writing, was carried out for Arabic-, Chinese-, and Spanish-speaking test takers. Two factors were identified, Speaking and a fusion of the other sections of the test. The number of factors, the factor loadings, and the factors' error variances were invariant in the three samples, although the correlations between the factors differed. The failure to find separate factors for each section of the LanguEdge test necessarily raises questions about the test's functioning that need to be resolved.

Key words: LanguEdge, English as a second language, factor analysis

Acknowledgements

Thanks are due to Kristin L. Huff and Cindy Nguyen for providing data and other information about the LanguEdge field test; Min Hwei Wang for doing the computer analysis; and Lyle F. Bachman, Michal Beller, Richard M. Luecht, and Yasuyo Sawaki for reviewing a draft of this report.

Table of Contents

Page

List of Tables

List of Figures

The recently introduced LanguEdge™ courseware (ETS, 2002) is intended to improve the learning of English as a second language (ESL) by providing classroom assessments of communicative skills. LanguEdge consists of two forms of a full-length, computer-administered linear ESL test (covering reading, writing, speaking, and listening) and supplementary material (teacher's guide, scoring handbook, and score interpretation guide). The test is broadly similar to the new version of the Test of English as a Foreign Language™ (TOEFL®) now under development, which also covers the same four skills and integrates, in a somewhat different format, speaking and writing with reading and listening (ETS, 2004). In contrast, the current TOEFL does not include speaking and uses completely separate reading, writing, and listening tasks (ETS, 2000a).

Useful information is already available about the psychometric characteristics of the LanguEdge test, including its correlations with a paper-based version of the TOEFL and ratings of English-language proficiency by the test takers and their instructors (ETS, 2002; Powers, Roever, Huff, & Trapani, 2003). However, other issues about the test's construct validity have not been addressed thus far: Are the four sections (Reading, Writing, Speaking, and Listening) measuring distinguishably different constructs, and are the same constructs being assessed in different language groups? The use of tasks that integrate different language skills while enhancing the validity of measurement of these skills may also blur the distinction between them. Nonetheless, the section scores, if they are to have diagnostic value, as intended, must provide some degree of different information about the test takers' performance.

Factor structure and its invariance across language groups are recurring issues about the TOEFL and other ESL tests, and have been extensively investigated. Factor analytic evidence about the number and nature of factors in ESL tests, much of it stemming from the controversy over whether there is a unitary language proficiency factor underlying language tasks (Oller, 1976), has been recently reviewed (Sasaki, 1999). The apparent consensus is that there is a general, higher-order factor as well as several narrower, first-order factors. Two studies are particularly relevant. Carroll (1983) reanalyzed a study by Scholz, Hendricks, Spurling, Johnson, and Vandenburg (1980) that included 22 variables: Comprehensive English Language Test (Harris & Palmer, 1970) Listening and Structure subtests, a modified version of the Reading for Understanding Placement Test (Thurstone, 1963), experimental oral interview scales, and experimental tests of listening, speaking, reading, writing, and grammar. An exploratory factor analysis by Carroll found five correlated first-order factors and a second-order, general factor. The first-order factors were Speaking, Listening, a fusion of Reading and Writing, oral comprehension and recall, and an uninterpretable factor. Bachman, Davidson, Ryan, and Choi (1995) analyzed a battery of 13 tests made up of the TOEFL (ETS, 1987; Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension sections), the Speaking Proficiency English Assessment Kit (ETS, 1985; the institutional version of the Test of Spoken English™; ETS, 1982), a test modeled after the Test of Written English™ (ETS, 1989), and the First Certificate in English (FCE; University of Cambridge Local Examinations Syndicate, 1987; Reading Comprehension, Composition, Use of English, Listening Comprehension, and Oral Interview subtests). An exploratory factor analysis found four correlated first-order factors and a second order, general factor. The first-order factors were Speaking, Listening, and two fusions of Reading and Writing. Kunnan (1995) subsequently reanalyzed the data for a portion of the sample in separate confirmatory factor analyses for two language groups Indo-European and non Indo-European. The same four first-order factors were identified.

The invariance of factors in ESL tests across language groups has not been investigated as thoroughly. Several studies are especially pertinent. Kunnan (1995), as already noted, found the same four correlated first-order factors in confirmatory factor analyses of Indo-European and non Indo-European language groups. Swinton and Powers (1980), in the first of three studies of invariance in the TOEFL, found consistent differences among language groups in two of the factors identified (the third factor, defined by Listening Comprehension, was common to all groups). For Spanish and German groups, the factors were consistent with the sections of the test: one factor combined Vocabulary and Reading Comprehension, and the other was a fusion of Structure and Written Expression. In contrast, for African, Arabic, Chinese, and Japanese groups, one factor combined Reading Comprehension with Structure and Written Expression; the other was Vocabulary. The two factors for a Farsi group were not readily interpretable. However, two subsequent studies by Hale et al. (1988) and Hale, Rock, and Jirele (1989) found the same two factors across language groups—one factor defined by Listening Comprehension and the second by the other sections of the test. The divergence between the Swinton and Powers results and those of Hale et al. in their two studies is probably attributable to differences in the factor analytic methods used: exploratory factor analyses of items for Swinton and Powers and confirmatory factor analyses of parcels of items for Hale et al. in their studies (Hale et al., 1989).

Accordingly, the aim of the present study was to assess the factors underlying the LanguEdge test and the invariance of these factors across language groups.

Method

Samples

The samples were drawn from test takers who participated in a field test of the LanguEdge test, taking the Form 1 of the test and a paper-based TOEFL, and completing self-ratings of English-language proficiency. The participants, paid volunteers from the TOEFL test-taking population, were recruited domestically and internationally, and tested at 18 domestic and 12 international test centers (ETS, 2002). The three largest language groups among the test takers, Arabic, Chinese, and Spanish, were used in the study. (These are also the largest groups taking the computer-adaptive TOEFL in 1999-2000 [ETS, 2000b].) Details about the samples in the study follow:

1. Arabic (*N* = 100). These are (a) 66 test takers with complete LanguEdge test data, including all 22 from the Cairo test center, for whom information on native language was unavailable (all had Arabic surnames); and (b) 34 with incomplete data limited to the Speaking section (18 missing 1 to 4 tasks, and 16 missing all 5 tasks), presumably because of technical problems in the test administration.
2. Chinese (*N* = 225). These include all 52 test takers in a Hong Kong test center, for whom information about native language was unavailable (all had Chinese surnames). The 225 test takers are all those with complete test data.
3. Spanish (*N* = 114). These are all test takers with complete test data.

The LanguEdge test scores and sex of the three samples are summarized in Table 1. The Chinese sample had an appreciably higher mean Reading score than the other samples, and the Arabic sample had an appreciably higher mean Speaking score than the others. The means for the other sections were similar for the three samples. The Arabic sample had an appreciably higher percentage of men than the other samples.

Table 1

LanguEdge Test Performance and Sex of the Samples

Variable	Arabic		Chinese		Spanish	
	N	Mean/ percent	*N*	Mean/ percent	*N*	Mean/ percent
Mean LanguEdge score						
Listening	100	17.53(5.63)	225	16.68(4.92)	114	15.60(5.15)
Reading	100	13.24(5.87)	225	16.00(5.20)	114	13.43(5.18)
Speaking	66	3.69(.74)	225	3.19(.67)	114	3.36(.74)
Writing	100	2.54(.88)	225	2.71(.80)	114	2.33(.76)
Percent male/female	100	72.0/38.0	224	41.1/58.9	106	49.5/50.5

Note. Standard deviations appear in parentheses.

Measures

The Listening section consists of six prompts, each with five to six multiple-choice items (most with a single correct answer, and a few "partial credit" items with more than one correct answer).[1] A total score over the six prompts is reported for the test. This is a scaled score that ranges from 1 to 25. 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 and the instability inherent in factor analyses of items (Gorsuch, 1983).

The Reading section consists of three passages, each with 12 to 13 multiple-choice items (most with a single correct answer and a few partial credit items with more than one correct answer).[2] A total score over the three passages is reported for the test. This is a scaled score that ranges from 1 to 25. 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 was used for the reasons already described.

The Speaking section consists of five speaking tasks, two of which (listening/speaking and reading/speaking) are administered as part of the Listening and Reading sections, respectively. Each task is rated on a 1 to 5 scale by experienced raters. The mean of these scores is reported for the test; it ranges from 1 to 5. For this study, the score for each task was obtained, and these five scores were used in the analysis.

The Writing section consists of three writing tasks, two of which (listening/writing and reading/writing) are administered as part of the Listening and Reading sections, respectively. Each task is rated on a 1 to 5 scale. The mean of these scores is reported for the test; it ranges from 1 to 5. For this study, the score for each task was obtained, and these three scores were used in the analysis.

Analysis

Missing scores on the Speaking section in the Arabic sample were imputed by the Markov chain Monte Carlo procedure (Tanner & Wong, 1987), with NORM (Schafer, 1999), using data for the entire Arabic sample (*N* = 165)—the five Speaking scores and three self-rating measures of speaking proficiency.[3] The three were a global rating ("Please rate your overall English language ability in … Speaking," with a five-point scale ranging from Extremely Good to Poor); a five-item "how well" scale (e.g., "I can speak for about one minute in response to a question," with a five-point scale ranging from Extremely Well to Not at All); and a five-item "agreement" scale (e.g., "I can state and support my opinion when I speak English," with a five-point scale ranging from Completely Agree to Completely Disagree). The imputed scores for the 18 test takers with partially missing Speaking scores and for a random sample of 16 out of 32 test takers with completely missing Speaking scores were used in the analysis (the number of imputed scores per task ranged from 22 to 30), along with the available Speaking scores for these 18 test takers and the 66 test takers with complete test data.

Multiple-group confirmatory factor analyses of the 17 scores for the three samples were conducted. The raw scores for the set of 17 variables in each sample were screened for multivariate outliers with the Mahalanobis distance function, using the .001 alpha level.[4] The scores for each of the samples were pooled and then normalized, using an area conversion of scores, with PRELIS 2 (Joreskog & Sorbom, 1996b). The normalized scores for each variable in each sample were evaluated for univariate skewness and kurtosis with standard tests of skewness and kurtosis, using the .001 familywise alpha level (Bonferroni adjustment). These scores for the set of 17 variables in each sample were also evaluated for multivariate kurtosis with the Mardia (1970) test, using the .001 alpha level. The scatterplots of these scores for all pairs of variables in each sample were inspected for nonlinearity. Covariance matrices for each sample were computed from the normalized scores and then analyzed by the 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. The models follow:

1. There are four correlated first-order factors corresponding to the test sections. (See Figure 1.) This model reflects the rationale underlying the test and is consistent with the familiar, four-fold categorization of the language domain along two independent dimensions: receptive vs. productive skills and oral vs. written language, with listening exemplifying a combination of a receptive skill and oral language, and so forth (Carroll, 1993). (In the event that this model is supported, an additional model would be tested: The four first-order factors are subsumed by a second-order, general factor.)
2. There is only one factor, made up of the four sections of the test. (See Figure 2.) This is an obvious model for a cognitive test.
3. There are two correlated factors, one for the Speaking section and one for the Listening, Reading, and Writing sections. (See Figure 3.) This model is based on the findings of an exploratory factor analysis of the test in a separate sample. (See Appendix A.)
4. There are three correlated factors, one for the Reading and Writing sections, one for the Listening section, and one for the Speaking section. (See Figure 4.) This model is based on the Carroll (1983), Bachman et al. (1995), and Kunnan (1995) findings.

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 models follow:

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.

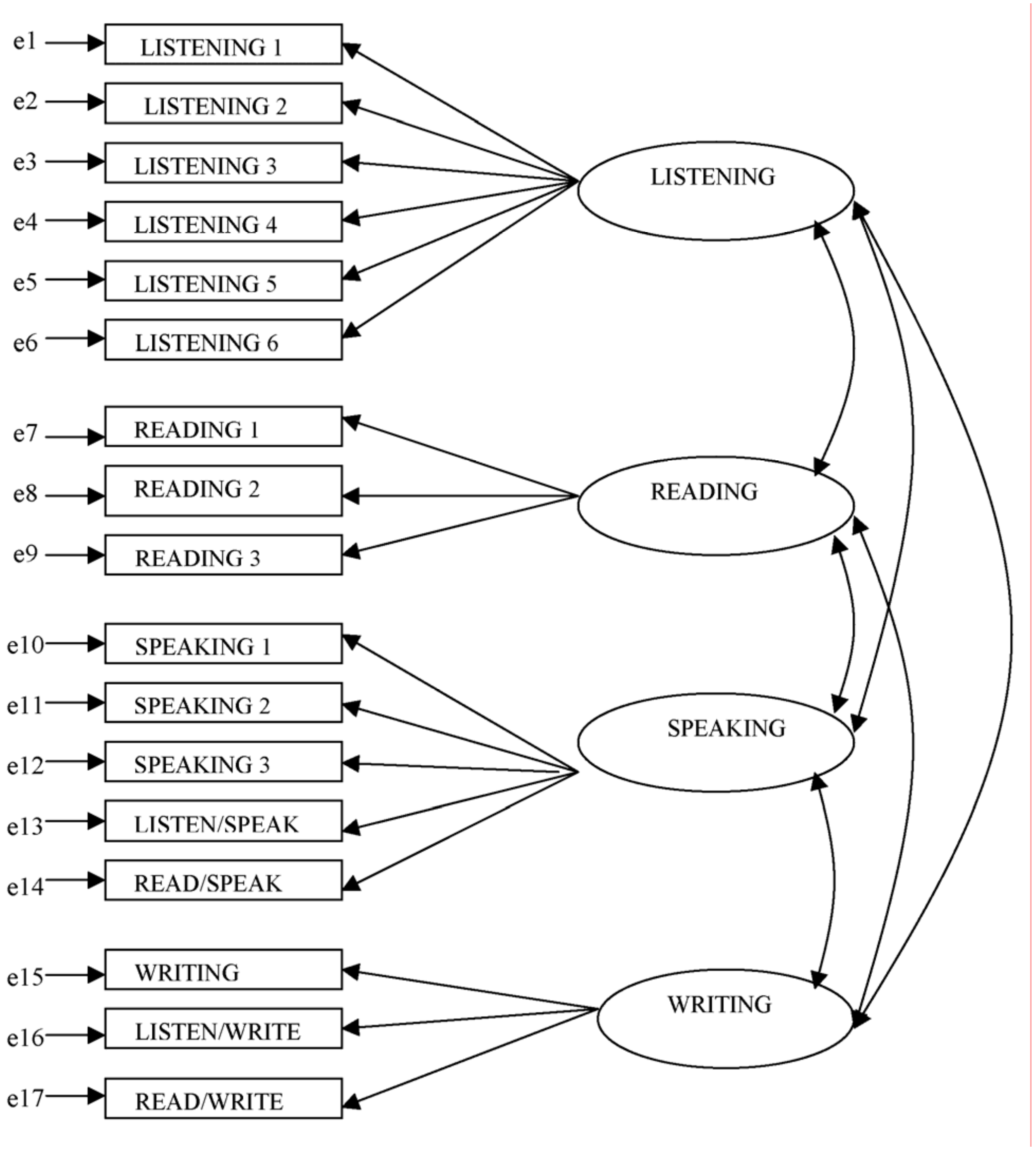

Figure 1. **Model 1: Four factors—Listening, Reading, Speaking, and Writing.**

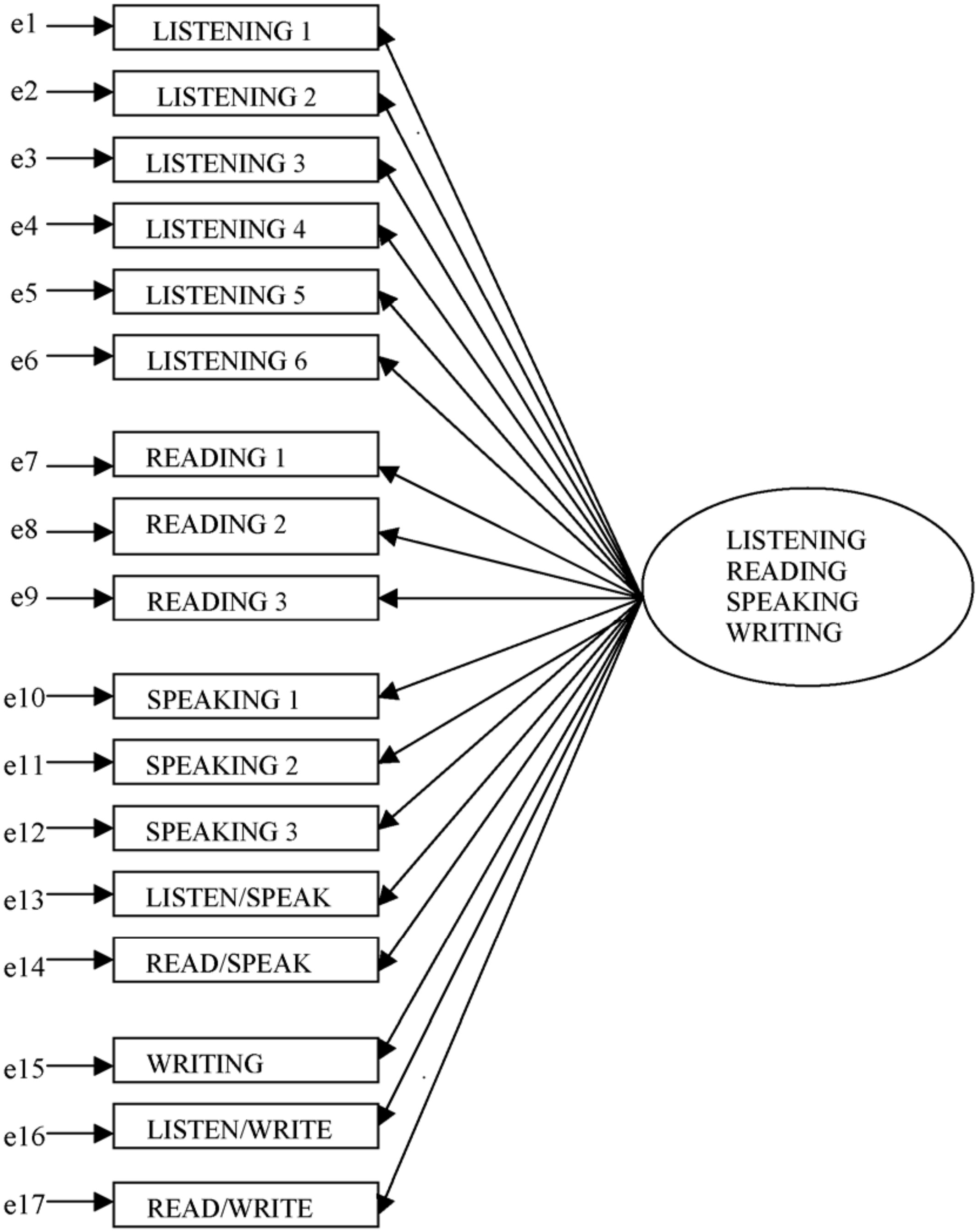

Figure 2. **Model 2: One factor—Listening, Reading, Speaking, and Writing.**

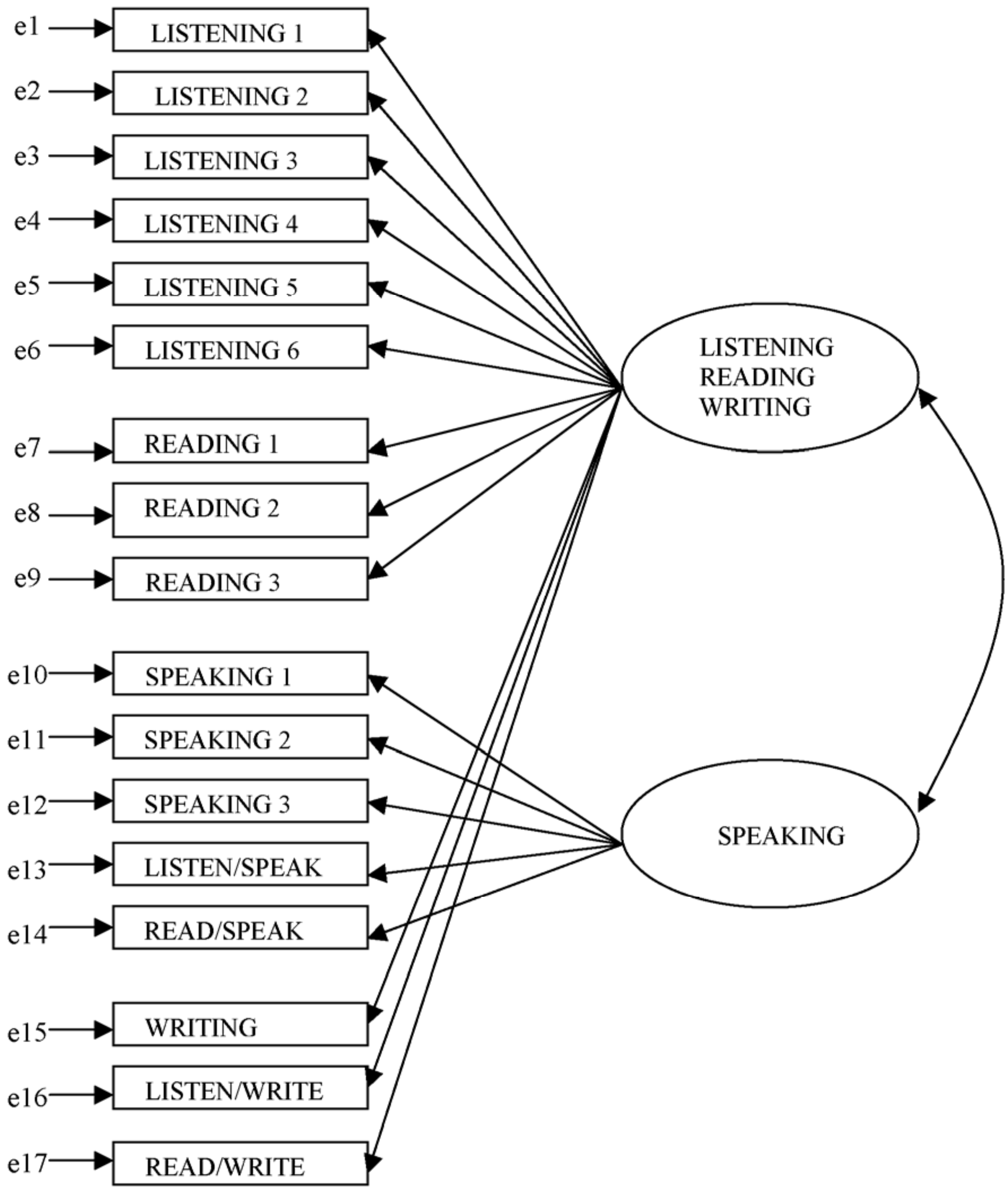

Figure 3. **Model 3: Two factors—Speaking vs. Listening, Reading, and Writing.**

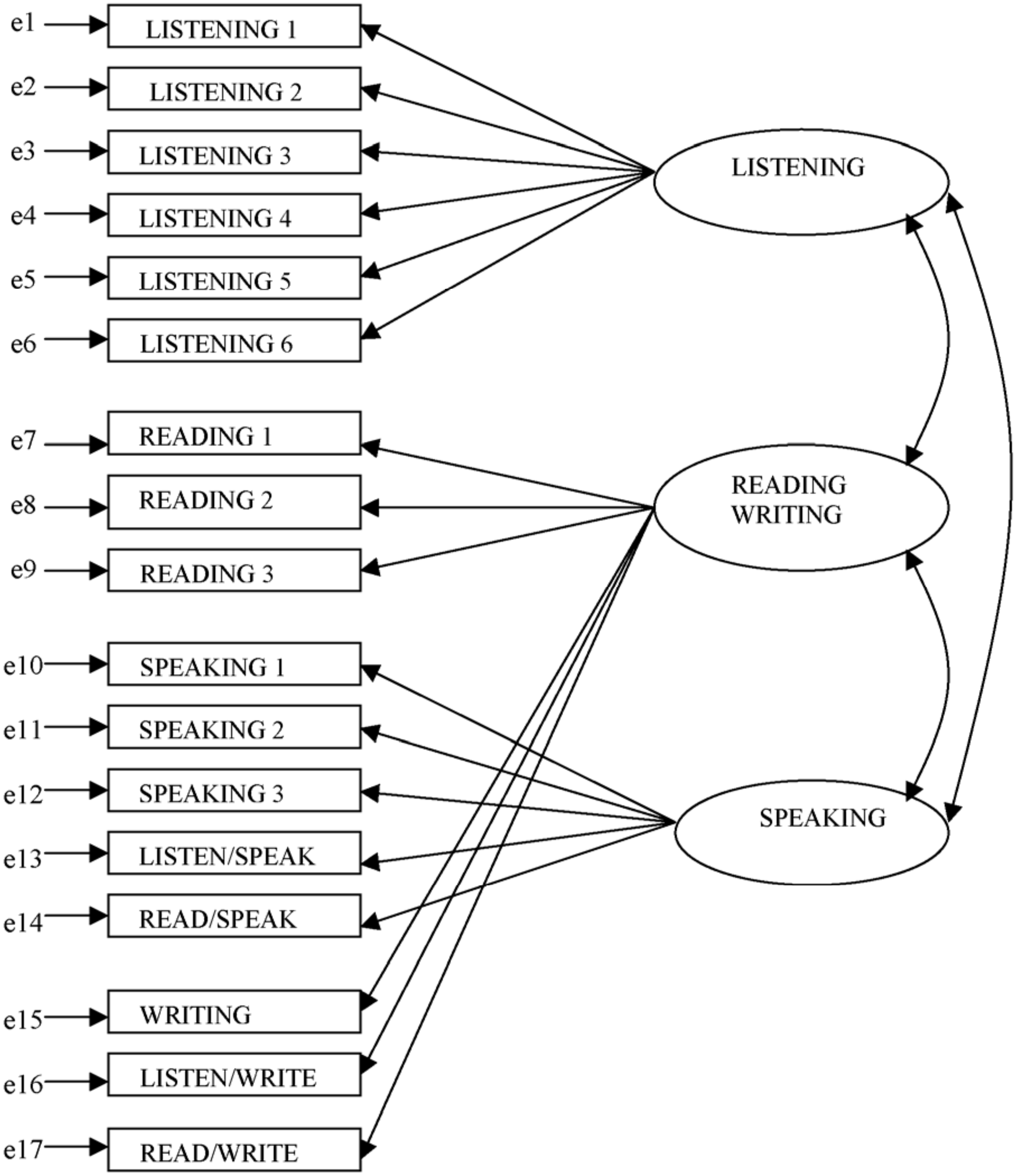

Figure 4. **Model 4: Three factors—Reading and Writing, Listening, and Speaking.**

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

Results and Discussion

Data Screening

There were no significant multivariate outliers for the raw scores of the variables in any sample. None of the corresponding normalized scores had significant univariate kurtosis in any sample. However, these scores for one variable, Reading: Passage 1, in the Chinese sample, had significant univariate skewness. None of the sets of these scores had significant multivariate kurtosis in any sample. And none of the scatterplots of these scores appeared nonlinear in any sample.

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. (The covariance matrices appear in Appendix B.)

For Model 1 (four factors), the goodness of fit indexes for the individual samples and for the overall analysis of the three samples were generally satisfactory. However, some of the correlations among the factors were extremely high, above .9 for one of the six correlations in the Arabic sample and for three of the correlations in each of the other samples, suggesting that this solution is implausible (Bagozzi & Yi, 1988).

For Model 2 (a single factor), most of the fit indexes for the individual samples and for the overall analysis were satisfactory, with the notable exceptions of the GFI of .79 for the Arabic sample and .84 for the Chinese sample, and the RMSEA of .08 for the overall analysis.

For Model 3 (two factors, one for Speaking and one for the three other sections), the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the important exception of the GFI of .85 for the Arabic sample, and there were no high correlations between the factors for any sample.

For Model 4 (three factors, one for Reading and Writing, one for Listening, and one for Speaking), the fit indexes for the individual samples and the overall analysis were generally satisfactory. But the high correlation between the Reading and Writing factor and the Listening factor in all of the samples suggests that this solution is implausible.

In short, only two of the models seem plausible, with reasonable but not perfect fits to the data: Model 2 (a single factor) and Model 3 (two factors, one for Speaking, one for the three other sections). Detailed comparisons were made of these two models. All of the χ^2 differences for the individual samples and the overall analyses were statistically ($p < .05$) and practically ($\chi^2/df > 3.00$) significant. The χ^2 differences was 41.92 (χ^2/df of 41.92) for the Arabic sample, 47.38 (χ^2/df of 47.38) for the Chinese sample, and 15.97 (χ^2/df of 15.97) for the Spanish sample, each with 1 *df*. The χ^2 difference was 105.26 for the overall analysis, with 3 *df* ($\chi^2/df = 35.09$). All of these differences reflect smaller χ^2s for Model 3 and indicate better fit for this solution. Similarly, the ECVIs were 2.08 for Model 2 and 1.71 for Model 3, and the respective 90% confidence intervals for the ECVIs were 1.92 to 2.06 and 1.58 to 1.87, indicating that Model 3 is expected to cross-validate better in a new sample. Based on these results, Model 3 was chosen for further analysis.

Model 3: Two Factors—Speaking and the Other Sections

The goodness of fit indexes and related information are reported in Table 3 for sequential tests about the invariance in factor loadings, error variances, and factor correlations for Model 3. (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 indexes of goodness of fit for the individual samples and for the overall analysis were generally satisfactory (the main exception is the GFI of .85 for the Arabic sample), and there were no high correlations between the factors. This outcome suggests that the number of factors is invariant.

Table 2

Tests of Invariance in Number of Factors: Four Models

Model and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Model 1: Four factors—Listening, Reading, Speaking, and Writing									
Arabic	113	142.10*	1.26	.05	.86				.93
Chinese	113	182.53**	1.62	.05	.91				.90, .91, .93
Spanish	113	112.71	1.00	.05	.90				.95, .97, .98
Overall	339	437.33**	1.29			.97	.99	.04	
Model 2: One factor—Listening, Reading, Speaking, and Writing									
Arabic	119	205.44**	1.73	.07	.79				--
Chinese	119	284.88**	2.39	.06	.84				--
Spanish	119	134.35	1.13	.05	.88				--
Overall	357	624.67**	1.75			.97	.97	.08	

(Table continues)

Table 2 (continued)

Model and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Model 3: Two factors—Speaking vs. Listening, Reading, and Writing									
Arabic	118	163.52**	1.39	.06	.85				None
Chinese	118	237.50**	2.01	.06	.88				None
Spanish	118	118.38	1.00	.05	.89				None
Overall	354	519.41**	1.47			.95	.94	.06	
Model 4: Three factors—Reading and Writing Listening, and Speaking									
Arabic	116	153.30**	1.32	.06	.86		.		.92
Chinese	116	233.29**	2.01	.06	.88				.96
Spanish	116	115.47	1.00	.05	.90				.97
Overall	348	502.05**	1.44			.99	.98	.06	

[a] Standardized root mean square residual. [b] Goodness of fit index. [c] Comparative fit index. [d] Nonnormed fit index. [e] Root mean square error of approximation.

[*] $p < .05$. [**] $p < .01$.

With regard to the invariance in the factor loadings across samples, again the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the important exception of the GFI of .84 for the Arabic sample, and there were no high correlations between the factors. Furthermore, the χ^2 difference between this overall analysis and the preceding overall analysis of the number of factors was not statistically significant ($p > .05$): 38.06, with 30 *df* ($\chi^2/df = 1.27$). These results suggest that the factor loadings are invariant.

With regard to the invariance in the factor loadings and error variances across samples, once again the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the notable exception of the GFI of .83 for the Arabic sample, and there were no high correlations between the factors. The χ^2 difference between this overall analysis and the preceding overall analysis of the factor loadings was statistically but not practically ($\chi^2/df < 3.00$) significant: 60.67, with 34 *df* ($\chi^2/df = 1.78$). These results suggest that the error variances as well as the factor loadings are invariant.

With regard to the invariance in the factor loadings, error variances, and factor correlations across samples, most of the fit indexes for the individual samples and for the overall analysis were satisfactory. Critical exceptions were the GFI of .82 and SRMR of .15 for the Arabic sample, which was accompanied by a positively skewed distribution of residuals, indicating that the hypothesized model underestimated the covariance matrix (Joreskog, 1993). The χ^2 difference between this overall analysis and the preceding overall analysis of factor loadings and error variances was statistically but not practically significant: 16.01, with 6 *df* ($\chi^2/df = 2.67$). The poor fit indexes for the Arabic sample suggest that the factor correlations are not invariant, though the factor loadings and error variances are invariant, as implied by the results for preceding models.

The model for invariant factor loadings and error variance is shown in Figure 5, with common metric, completely standardized factor loadings and error variances. The factor loadings were substantial (.5 or more). The correlation between the factors was somewhat lower for the Arabic sample (.73) than for the Chinese and Spanish samples (.83 and .86, respectively), a practically significant, “medium” level difference (Cohen, 1988).

Table 3

Complete Tests of Invariance in Factors, Model 3: Two Factors—Speaking vs. Listening, Reading, and Writing

Hypothesis and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Number of factors invariant									
Arabic				.06	.85				None
Chinese				.06	.88				None
Spanish				.05	.89				None
Overall	354	519.41**	1.47			.95	.94	.06	
Factor loadings invariant									
Arabic				.07	.84				None
Chinese				.07	.87				None
Spanish				.09	.88				None
Overall	384	557.47**	1.45			.95	.94	.06	

(Table continues)

Table 3 (continued)

Hypothesis and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Factor loadings and error variances invariant									
Arabic				.08	.83				None
Chinese				.08	.86				None
Spanish				.09	.87				None
Overall	418	618.14**	1.48			.94	.94	.06	
Factor loadings, error variances, and factor correlations invariant									
Arabic				.15	.82				
Chinese				.09	.86				
Spanish				.09	.87				
Overall	424	634.15**	1.50			.94	.94	.06	

[a] Standardized root mean square residual. [b] Goodness of fit index. [c] Comparative fit index. [d] Nonnormed fit index. [e] Root mean square error of approximation.

[*] $p < .05$. [**] $p < .01$.

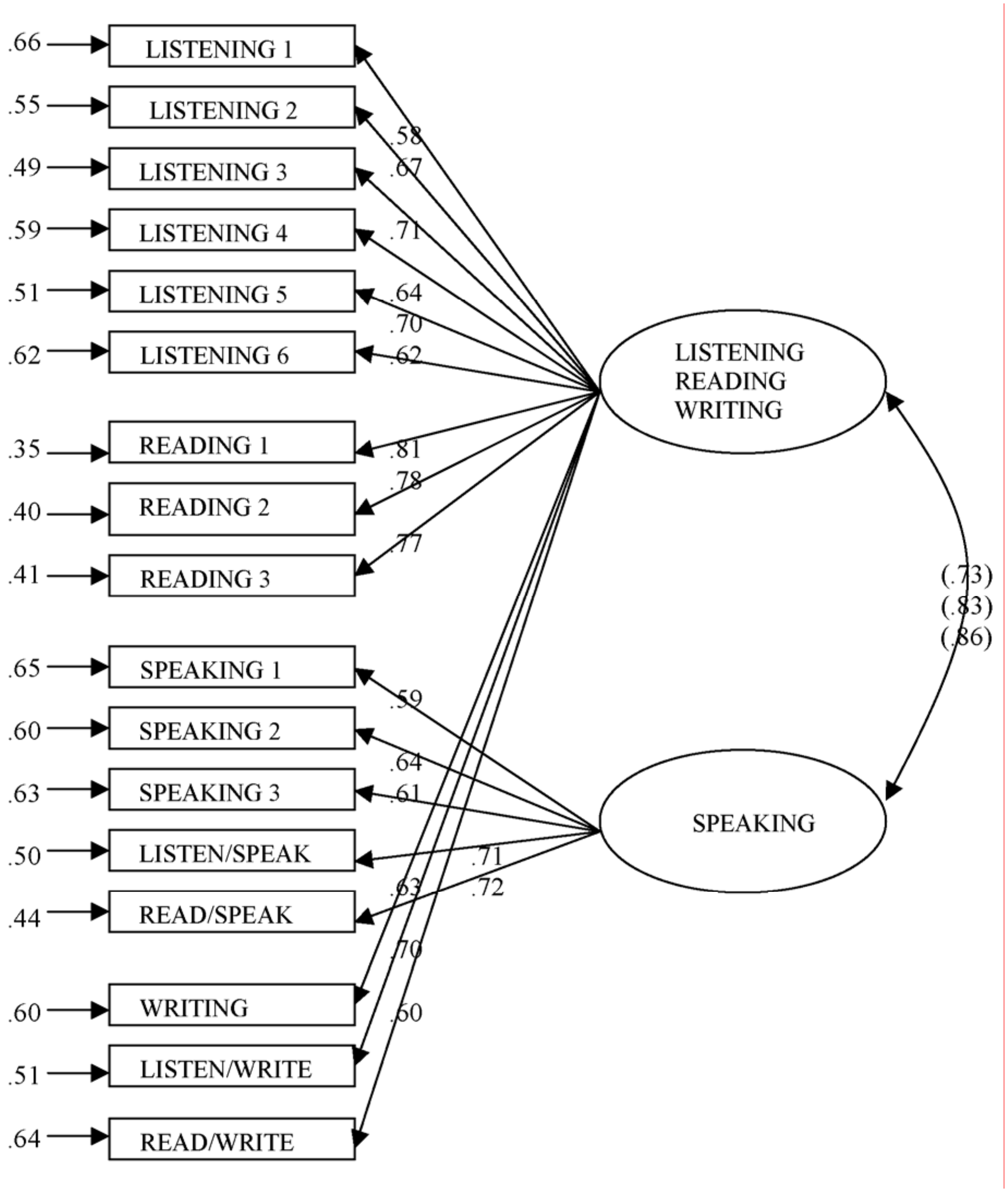

Figure 5. **Model 3: Two factors—Speaking vs. Listening, Reading, and Writing, with common metric, completely standardized factor loadings, and error variances shown.**

Conclusions

It is important not to overinterpret the results of this study. They are based on only three language groups and relatively modest samples. It is uncertain whether the samples, though recruited from the TOEFL test taking population, are representative. And the fit between the data and the models was far from ideal. Further hypothesis-testing studies with more and larger representative samples of language groups are needed to assess the generalizability of the unexpected outcomes that were obtained.

A key finding was that the four sections of the LanguEdge test seem to represent two distinct but correlated factors, Speaking, and a fusion of Listening, Reading, and Writing, not four factors corresponding to the sections of the test. This outcome is consistent with the results for the exploratory factor analysis of the test, which is noteworthy because of the great diversity of language groups in the composite sample of test takers used in that analysis. However, this finding diverges from the three factors (one for Reading and Writing, and one factor each for Listening and Speaking) identified by Carroll (1983), Bachman et al. (1995), and Kunnan (1995). This difference may be attributable to limitations in the previous studies. The Scholz et al. (1980) study reanalyzed by Carroll had considerable missing data and small *N*s for individual correlations (pair-wise correlations were used). *N*s ranged from 65 to 162 for a total sample of 186. The resulting sampling error in the correlations and a non-Gramian correlation matrix may have distorted the factors obtained. And Kunnan found a very high correlation (.92) between two of the Bachman et al. factors for non Indo-European test takers in his confirmatory factor analyses of the Bachman et al. data for part of the original sample, suggesting that the four-factor solution identified by him in these analyses and previously by Bachman et al. in their exploratory factor analysis may be implausible.

Perhaps more important, the combined listening, reading, and writing factor obtained in the present study disagrees with the findings of a number of investigations that found a listening factor that is separate from a reading factor. They include not only the Carroll (1983), Bachman et al. (1995), and Kunnan (1995) studies, but also studies of the TOEFL by Swinton and Powers (1980) and Hale et al. (1988, 1989), cited earlier, and by Manning (1987), an exploratory factor analysis of item parcels. (See also the review by Buck, 1992.) The reason for this divergence is unclear. It does not appear to be attributable to an unusual feature of the LanguEdge test: the presence of some integrative tasks (listening/speaking, reading/speaking, listening/writing, reading/writing) that may tap more than one skill. None of the tasks involved both listening and reading, and the results were similar in secondary analyses that permitted the four integrative tasks to define both the factor for their own test section and the factor for the test section in which they were administered. (See Appendix C.)

One speculation is that the clear distinction between speaking and other language skills in this study reflects the effect of ESL instructional practices and, indirectly, the influence of the TOEFL on this instruction (e.g., Alderson & Wall, 1993). The TOEFL covers listening, reading, and writing, but not speaking, and the use of the separate TSE is largely limited to test takers planning to be teaching assistants. Hence, given the popularity of the TOEFL, ESL instruction may de-emphasize speaking while producing some degree of uniformity in other language skills, with the consequence that measures of these skills are highly related to each other but less related to speaking measures.

The other central finding was the largely invariant factor structure across the language groups (differences in factor correlations were an exception), indicating that the LanguEdge test was operating similarly for these groups. This outcome is broadly congruent with the results by Hale et al. (1988, 1989), who observed invariance in TOEFL factors, and by Kunnan (1995), who found invariance in TOEFL and FCE factors. (Although these studies used the same confirmatory factor analytic methods as the present one, they did not make analytical comparisons of the invariance of the factor loadings, error variances, and factor correlations.) The general invariance in this study and its absence in the Swinton and Powers (1980) study of the TOEFL, like the divergence between their findings and those in the Hale et al. studies, are likely due to differences in analytical methods. The invariance in this study is also consistent with Brown’s (1999) finding in a generalizability study of the TOEFL that language group accounted for a minimal amount of test score variance relative to the variance associated with persons, items, and subtests.

The study’s failure to find separate factors for each section of the LanguEdge test necessarily raises questions about the test’s functioning that need to be resolved. The Hale et al. (1989) observation in connection with their factor analyses of the TOEFL bears repeating: whether or not the sections of the LanguEdge test turn out to be empirically distinguishable, and hence practically useful for differential diagnosis of ESL competency, has no necessary bearing on the theoretical value of the conceptual distinctions between the reading, writing, listening, and speaking skills that these test sections assess.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14,* 115–129.

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.

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

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

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16,* 217–238.

Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning, 42,* 313-357.

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.

Carroll, J. B. (1985). Exploratory factor analysis: A tutorial. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. 1. Research methodology* (pp. 25–58). Norwood, NJ: Ablex.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

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

ETS. (1982). *Test of Spoken English Manual for score users* (Rev. ed.). Princeton, NJ: Author.

ETS. (1985). *Guide to SPEAK.* Princeton, NJ: Author.

ETS. (1987). *TOEFL test and score manual.* Princeton, NJ: Author.

ETS. (1989). *TOEFL Test of Written English guide.* Princeton, NJ: Author.

ETS. (2000a). *Computer-based TOEFL score user guide* (2000–2001 ed.) Princeton, NJ: Author.

ETS. (2000b). *TOEFL test and score data summary* (2000–2001 ed.). Princeton, NJ: Author.

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. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Rep. No. RR-32; ETS RR-89-42). Princeton, NJ: ETS.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W., Jr. (1988). *Multiple-choice cloze items and the Test of English as a Foreign Language* (TOEFL Research Rep. No. RR-26; ETS RR-88-2). Princeton, NJ: ETS.

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimum residuals (minres). *Psychometrika, 31,* 351–368.

Harris, D. P., & Palmer, L. A. (1970). *CELT technical manual for Listening Form L-A, Structure Form S-A, Vocabulary Form V-A.* New York: McGraw-Hill.

Hendrickson, A. E., & White, P.O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17,* 65–70.

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.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31,* 313–323.

Joreskog, K. G. (1993). Testing structural models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, 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.

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.

Manning, W. H. (1987). *Development of cloze-elide tests of English as a second language* (TOEFL Research Rep. No. RR-23; ETS RR-87-18). Princeton, NJ: ETS.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

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

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge courseware scores against faculty ratings and student self-assessments* (ETS RR-03-11). Princeton, NJ: ETS.

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.

Schafer, J. L. (1999). NORM: Multiple imputations of incomplete multivariate data under a normal model (Version 2.03 software for Windows 95/98/NT) [Computer software and manual]. Retrieved April 27, 2004, from http://www.stat.psu.edu/~jls/misoftwa.html

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? A factor analysis of twenty-two English proficiency tests. In J. W. Oller, Jr., & K. Perkins (Ed.), *Research in language testing* (pp. 24-33). Rowley, MA: Newbury House.

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

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Rep. No. RR-06; ETS RR-80-32). Princeton, NJ: ETS.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82,* 528-550.

Thurstone, T. G. (1963). *Manual, Reading for Understanding Placement Test*. Chicago: Science Research Associates.

University of Cambridge Local Examinations Syndicate. (1987). *English as a Foreign Language: General handbook.* Cambridge, England: Author.

Notes

[1] Standard items earn one point if answered correctly; partial credit items earn one point if more than one of the answers is correct.

[2] Standard items earn one point if answered correctly; partial credit items earn two or three points if more than one of the answers is correct.

[3] For the 114 test takers with at least one Speaking score, the multiple correlations of the three self-ratings with the Speaking scores ranged from .33 to .47; the multiple correlations of the self-ratings and each set of four Speaking scores with the fifth Speaking score ranged from .60 to .70.

[4] This analysis for the Arabic sample was limited to the 66 test takers with complete LanguEdge scores.

Appendix A

Exploratory Factor Analysis

An exploratory, principal factor analysis was carried out for the 436 test takers from the 47 language groups in the field test not used in the main analysis. The LanguEdge test scores and sex of the sample are summarized in Table A1. The product-moment intercorrelations of the 17 LanguEdge scores, reported in Table A2, were analyzed by the minres method (Harman & Jones, 1966). The number of factors was determined on the basis of a scree test of eigenvalues (Cattell, 1966) and the salient rotated factors loadings (pattern coefficients) in analyses done for varying numbers of factors (Carroll, 1985). Oblique rotations of the factors were carried out with the promax (Hendrickson & White, 1964) and oblimin (Jennrich & Sampson, 1966) methods.

The different methods varied in the number of factors that they identified: either two or three with the scree test, two to five with salients in a promax rotation, and either two or four with salients in an oblimin rotation. Two factors were chosen from a consensus of these results.

The two factors correlated .71 with the promax rotation and .65 with the oblimin rotation. The rotated loadings for the two kinds of rotations are reported in Table A3. The pattern of salient loadings for the two rotation methods is identical. Factor I was consistently defined by most of the Listening, Reading, and Writing scores, and the Reading/Speaking score. Factor II was consistently defined by all of the Speaking scores except the Reading/Speaking score.

Table A1

LanguEdge Test Performance and Sex of the Composite Sample

Variable	*N*	Mean/ percent
Mean LanguEdge score		
Listening	436	17.68(5.24)
Reading	436	16.27(5.38)
Speaking	436	3.41(.81)
Writing	436	2.83(.90)
Percent male/female	433	52.2/47.8

Note. Standard deviations appear in parentheses.

Table A2

Correlation Matrix for Composite Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Listening—Prompt 1	1.00																
2. Listening—Prompt 2	.43	1.00															
3. Listening—Prompt 3	.44	.59	1.00														
4. Listening—Prompt 4	.40	.48	.48	1.00													
5. Listening—Prompt 5	.49	.50	.58	.48	1.00												
6. Listening—Prompt 6	.37	.51	.50	.54	.53	1.00											
7. Reading—Passage 1	.42	.59	.62	.56	.52	.57	1.00										
8. Reading—Passage 2	.39	.54	.50	.52	.44	.54	.70	1.00									
9. Reading—Passage 3	.42	.55	.55	.56	.48	.55	.70	.67	1.00								
10. Listening/Speaking	.39	.46	.57	.37	.50	.44	.48	.41	.43	1.00							
11. Reading/Speaking	.43	.52	.56	.48	.51	.46	.63	.58	.57	.49	1.00						
12. Speaking—Task 1	.33	.29	.42	.27	.36	.32	.35	.28	.23	.48	.39	1.00					
13. Speaking—Task 2	.33	.38	.45	.31	.45	.36	.40	.39	.28	.50	.49	.50	1.00				
14. Speaking—Task 3	.43	.42	.47	.42	.49	.39	.41	.37	.38	.55	.51	.48	.51	1.00			
15. Listening/Writing	.38	.50	.53	.49	.47	.49	.56	.53	.54	.50	.55	.32	.40	.41	1.00		
16. Reading/Writing	.31	.42	.44	.42	.41	.46	.55	.53	.51	.40	.49	.28	.32	.33	.54	1.00	
17. Writing	.39	.44	.48	.42	.41	.42	.54	.49	.50	.43	.50	.34	.38	.38	.58	.51	1.00

Table A3

Factor Loadings for Composite Sample

	Oblimin rotation		Promax rotation		
Variable	Factor I	Factor II	Factor I	Factor II	h^2
Listening—Prompt 1	*.34*	*.30*	*.31*	*.32*	.66
Listening—Prompt 2	*.62*	.13	*.60*	.15	.50
Listening—Prompt 3	*.50*	*.34*	*.45*	*.37*	.43
Listening—Prompt 4	*.66*	.03	*.66*	.04	.53
Listening—Prompt 5	*.41*	*.37*	*.36*	*.41*	.50
Listening—Prompt 6	*.64*	.09	*.62*	.10	.52
Reading—Passage 1	*.87*	-.04	*.87*	-.03	.29
Reading—Passage 2	*.85*	-.11	*.86*	-.11	.38
Reading—Passage 3	*.95*	-.21	*.97*	-.22	.31
Speaking—Task 1	-.07	*.71*	-.16	*.77*	.55
Speaking—Task 2	.03	*.68*	-.05	*.74*	.50
Speaking—Task 3	.09	*.67*	.01	*.72*	.46
Listening/Speaking	.20	*.59*	.13	*.64*	.45
Reading/Speaking	*.55*	.27	*.51*	*.30*	.43
Writing	*.56*	.14	*.54*	.16	.56
Listening/Writing	*.62*	.14	*.60*	.16	.48
Reading/Writing	*.66*	.00	*.65*	.01	.57

Note. The factor loadings reported are pattern coefficients. Loadings of ± .30 or above are in italics.

Appendix B

Covariance Matrices for Arabic, Chinese, and Spanish Samples

Table B1

Covariance Matrix for Arabic Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Listening—Prompt 1	.87																
2. Listening—Prompt 2	.59	1.11															
3. Listening—Prompt 3	.73	1.01	3.02														
4. Listening—Prompt 4	.57	.86	1.32	2.42													
5. Listening—Prompt 5	.61	.88	1.49	1.02	2.42												
6. Listening—Prompt 6	.51	.96	1.55	1.33	1.37	3.08											
7. Reading—Passage 1	1.38	1.75	3.58	2.73	3.27	3.18	10.61										
8. Reading—Passage 2	1.25	1.55	3.10	2.44	2.81	3.12	7.32	10.24									
9. Reading—Passage 3	1.59	2.16	3.40	2.45	3.18	3.36	7.89	7.35	11.69								
10. Listening/Speaking	.54	.51	.96	.80	.82	.75	2.13	1.60	2.07	1.64							
11. Reading/Speaking	.47	.46	.64	.91	.47	.76	1.83	1.46	1.59	.88	1.63						
12. Speaking—Task 1	.38	.35	.70	.41	.75	.31	1.36	1.07	1.22	.50	.59	1.03					
13. Speaking—Task 2	.44	.42	.63	.60	.62	.51	1.49	1.04	1.47	.69	.66	.48	1.16				
14. Speaking—Task 3	.21	.28	.58	.44	.56	.46	1.17	.88	1.14	.49	.46	.40	.42	.78			
15. Listening/Writing	.47	.57	1.15	.75	.88	1.03	2.23	2.09	2.16	.68	.68	.49	.48	.41	1.16		
16. Reading/Writing	.35	.42	.68	.53	.69	.61	1.86	1.88	2.22	.54	.45	.23	.22	.14	.59	1.01	
17. Writing	.36	.44	.93	.65	.83	.50	1.81	1.59	1.95	.58	.44	.46	.48	.39	.76	.55	1.21

Table B2

Covariance Matrix for Chinese Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Listening—Prompt 1	.89																
2. Listening—Prompt 2	.30	.95															
3. Listening—Prompt 3	.43	.69	2.38														
4. Listening—Prompt 4	.55	.62	1.13	2.39													
5. Listening—Prompt 5	.47	.71	1.07	1.31	2.28												
6. Listening—Prompt 6	.43	.56	.92	.96	.88	2.01											
7. Reading—Passage 1	1.14	1.20	2.67	2.27	1.88	1.85	7.61										
8. Reading—Passage 2	.96	1.18	2.34	2.18	1.89	1.85	4.85	7.25									
9. Reading—Passage 3	1.44	1.47	2.25	2.47	2.11	2.34	6.09	6.04	11.29								
10. Listening/Speaking	.33	.51	1.05	.91	.91	.53	1.59	1.42	1.46	1.36							
11. Reading/Speaking	.36	.42	.70	.67	.73	.58	1.51	1.30	1.72	.60	1.01						
12. Speaking—Task 1	.16	.18	.28	.36	.42	.24	.46	.52	.61	.40	.30	.58					
13. Speaking—Task 2	.12	.21	.36	.31	.36	.20	.59	.58	.68	.39	.36	.24	.60				
14. Speaking—Task 3	.07	.18	.37	.35	.42	.18	.56	.64	.48	.38	.29	.22	.27	.62			
15. Listening/Writing	.25	.44	.68	.73	.66	.52	1.27	1.21	1.23	.56	.46	.19	.26	.30	.87		
16. Reading/Writing	.27	.40	.71	.63	.57	.55	1.16	1.48	1.60	.49	.42	.16	.25	.21	.42	1.10	
17. Writing	.31	.42	.67	.70	.71	.48	1.19	1.12	1.50	.59	.55	.28	.29	.29	.50	.42	1.09

Table B3

Covariance Matrix for Spanish Sample

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1. Listening—Prompt 1	1.11																
2. Listening—Prompt 2	.52	.97															
3. Listening—Prompt 3	.84	.91	2.90														
4. Listening—Prompt 4	.55	.49	.83	1.76													
5. Listening—Prompt 5	.81	.77	1.56	.87	2.12												
6. Listening—Prompt 6	.56	.48	1.08	.84	.94	2.17											
7. Reading—Passage 1	1.60	1.75	2.74	1.61	2.54	2.12	7.97										
8. Reading—Passage 2	1.46	1.36	2.40	1.66	2.04	1.96	4.79	7.50									
9. Reading—Passage 3	1.92	1.70	3.04	1.86	2.37	2.12	5.56	5.61	9.33								
10. Listening/Speaking	.55	.59	.85	.51	.76	.40	1.38	1.20	1.30	1.45							
11. Reading/Speaking	.56	.54	.90	.63	.85	.68	1.78	1.66	1.76	.51	1.44						
12. Speaking—Task 1	.34	.30	.52	.35	.62	.34	.94	.81	.83	.45	.41	.77					
13. Speaking—Task 2	.35	.33	.57	.31	.49	.34	.95	.96	.98	.52	.41	.26	.76				
14. Speaking—Task 3	.41	.34	.67	.40	.66	.44	1.30	1.20	1.25	.45	.46	.37	.39	.81			
15. Listening/Writing	.50	.46	.63	.50	.76	.60	1.48	1.39	1.36	.35	.51	.26	.32	.40	.91		
16. Reading/Writing	.31	.36	.52	.27	.51	.44	1.12	1.20	1.29	.31	.34	.22	.19	.30	.27	.80	
17. Writing	.64	.45	.81	.33	.78	.65	1.60	1.63	1.83	.37	.44	.39	.49	.50	.48	.43	1.25

Appendix C

Confirmatory Factor Analyses of Alternative Models

Alternative Versions of Three Models

Confirmatory factor analyses were conducted, using the same methods in the main analyses, for alternative versions of Models 1, 3, and 4 that permitted the two Speaking tasks and the two Writing tasks administered as part of the Listening and Reading sections to define not only the same factor as the other tasks in their own test section but also the factor for the tasks in the test sections in which they were administered. (See Figures C1 to C3 for alternative Models 1, 3, and 4, respectively, and Figure C4 for Model 2, a single factor, which did not require reanalysis.) The goodness of fit indexes and related information are reported in Table C1, for the three models about the number of factors, as well as for Model 2.

For Model 1 (four factors), the goodness of fit indexes for the individual samples and for the overall analysis of the three samples were generally satisfactory. However, some of the correlations among the factors were extremely high, above .9 for one of the six correlations in the Arabic sample and for one of the correlations in the Spanish sample, suggesting that this solution is implausible.

For Model 2 (a single factor), as previously reported, most of the fit indexes for the individual samples and for the overall analysis were satisfactory, with the notable exceptions of the GFI of .79 for the Arabic sample and .84 for the Chinese sample, and the RMSEA of .08 for the overall analysis.

For Model 3 (two factors, one for Speaking and one for the three other sections), the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the important exception of the GFI of .85 for the Arabic sample, and there were no high correlations between the factors for any sample.

For Model 4 (three factors, one for Reading and Writing, one for Listening, and one for Speaking), the fit indexes for the individual samples and the overall analysis were generally satisfactory. However, the high correlation between the Reading and Writing factor and the Listening factor in all of the samples, as well as a factor loading over one for the Chinese sample, suggests that this solution is implausible.

In short, two models seem plausible, with reasonable fits to the data: Model 2 (a single factor) and Model 3 (two factors, one for Speaking, one for the three other sections). Detailed comparisons were made of these two models. All of the χ^2 differences for the individual samples and the overall analyses were statistically ($p < .05$) and practically ($\chi^2/df > 3.00$) significant. The χ^2 differences were 43.13 (χ^2/df of 14.38) for the Arabic sample, 79.04 (χ^2/df of 26.35) for the Chinese sample, and 21.05 (χ^2/df of 7.02) for the Spanish sample, each with 3 *df*. The χ^2 difference was 143.22 for the overall analysis, with 9 *df* ($\chi^2/df = 15.91$). All of these differences reflect smaller χ^2s for Model 3 and indicate better fit for this solution. Similarly, the ECVIs were 2.08 for Model 2 and 1.63 for Model 3, and the respective 90% confidence intervals for the ECVIs were 1.92 to 2.06 and 1.51 to 1.78, indicating that Model 3 is expected to cross-validate better in a new sample. Based on these results, Model 3 was chosen for further analysis.

Alternative Version of Model 3: Two Factors—Speaking and the Other Sections

Further analyses were conducted for the alternative version of Model 3. The goodness of fit indexes and related information are reported in Table C2 for sequential tests about the invariance in factor loadings, error variances, and factor correlations for this model. (Data for the invariance of the number of factors, previously described and reported, are repeated.)

With regard to the invariance in the number of factors across samples, as already noted, the goodness of fit indexes for the individual samples and for the overall analysis were generally satisfactory (an important exception was the GFI of .85 for the Arabic sample), and there were no high correlations between the factors. This outcome suggests that the number of factors is invariant.

With regard to the invariance in the factor loadings across samples, again the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the notable exception of the GFI of .84 for the Arabic sample, and there were no high correlations between the factors. Furthermore, the χ^2 difference between this overall analysis and the preceding overall analysis of the number of factors was not statistically significant ($p < .05$): 48.06, with 34 *df* ($\chi^2/df = 1.41$). These results suggest that the factor loadings are invariant.

Table C1

Tests of Invariance in Number of Factors: Alternative Models

Model and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Model 1: Four factors—Listening, Reading, Speaking, and Writing									
Arabic									
Chinese	109	131.03	1.20	.05	.87				.91
Spanish	109	138.80*	1.27	.04	.93				None
Overall	109	102.26	.94	.04	.91				.96
Model 2: One factor—Listening, Reading, Speaking, and Writing									
Arabic	119	205.44**	1.73	.07	.79				--
Chinese	119	284.88**	2.39	.06	.84				--
Spanish	119	134.35	1.13	.05	.88				--
Overall	357	624.67**	1.75			.97	.97	.08	

(Table continues)

Table C1 (continued)

Model and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Model 3: Two factors—Speaking vs. Listening, Reading, and Writing									
Arabic	116	162.31**	1.40	.06	.85				None
Chinese	116	205.84**	1.77	.05	.89				None
Spanish	116	113.30	.98	.04	.90				None
Overall	348	481.45**	1.38			.99	.99	.05	
Model 4: Three factors—Reading and Writing, Listening, and Speaking									
Arabic	113	150.85*	1.33	.06	.86		.		.91
Chinese	113	183.95**	1.63	.05	.91				.94
Spanish	113	108.73	.96	.04	.90				.96
Overall	339	443.53**	1.31			.99	.99	.04	

[a] Standardized root mean square residual. [b] Goodness of fit index. [c] Comparative fit index. [d] Nonnormed fit index. [e] Root mean square error of approximation.

$p < .05$. $^{**}p < .01$.

With regard to the invariance in the factor loadings and error variances across samples, once again the fit indexes for the individual samples and for the overall analysis were generally satisfactory, with the important exception of the GFI of .82 for the Arabic sample, and there were no high correlations between the factors. The χ^2 difference between this overall analysis and the preceding overall analysis of the factor loadings was statistically but not practically ($\chi^2/df >$ 3.00) significant: 69.39, with 34 *df* ($\chi^2/df = 2.04$). These results suggest that the error variances as well as the factor loadings are invariant.

With regard to the invariance in the factor loadings, error variances, and factor correlations across samples, most of the fit indexes for the individual samples and for the overall analysis were satisfactory. Critical exceptions were the GFI of .82 and SRMR of .15 for the Arabic sample. In addition, the distribution of residuals was skewed for this sample and for the Chinese sample, positively for the former and negatively for the latter, indicating that the hypothesized model underestimated the covariance matrix for the Arabic sample and overestimated it for the Chinese sample. The χ^2 difference between this overall analysis and the preceding, overall analysis of factor loadings and error variances was statistically, but not practically significant: 14.66, with 6 *df* (χ^2/df ratio of 2.44). The poor fit results for the Arabic and Chinese samples suggest that the factor correlations are not invariant, though the factor loadings and error variances are invariant, as implied by the results for preceding models.

The model for invariant factor loadings and error variances, with common metric, completely standardized factor loadings and error variances is shown in Figure C-5. The factor loadings were generally substantial. The two Speaking tasks permitted to define both factors were an exception. These tasks had relatively appreciable loadings (.44 to .46) on the Speaking factor and noticeably lower loadings on the combined Listening, Reading, and Writing factor. The correlation between the factors was somewhat higher for the Spanish sample (.80) than for the Arabic and Chinese samples (.66, and .69, respectively), a practically significant, “medium” difference.

Table C2

Complete Tests of Invariance in Factors, Alternative Model 3: Two Factors—Speaking vs. Listening, Reading, and Writing

Hypothesis and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Number of factors invariant									
Arabic				.06	.85				None
Chinese				.05	.89				None
Spanish				.04	.90				None
Overall	348	481.45**	1.38			.99	.99	.05	
Factor loadings invariant									
Arabic				.08	.84				None
Chinese				.06	.89				None
Spanish				.09	.88				None
Overall	382	529.51**	1.39			.99	.99	.05	

(Table continues)

Table C2 (continued)

Hypothesis and sample	*df*	χ^2	χ^2/df	SRMR[a]	GFI[b]	CFI[c]	NNFI[d]	RMSEA[e]	Factor correlations exceeding .9
Factor loadings and error variances invariant									
Arabic				.08	.82				None
Chinese				.07	.87				None
Spanish				.09	.87				None
Overall	416	598.90**	1.44			.98	.98	.06	
Factor loadings, error variances, and factor correlations invariant									
Arabic				.15	.82				
Chinese				.08	.87				
Spanish				.10	.87				
Overall	422	613.56**	1.45			.98	.98	.06	

[a] Standardized root mean square residual. [b] Goodness of fit index. [c] Comparative fit index. [d] Nonnormed fit index. [e] Root mean square error of approximation.

* *p* < .05. *p* < .01.

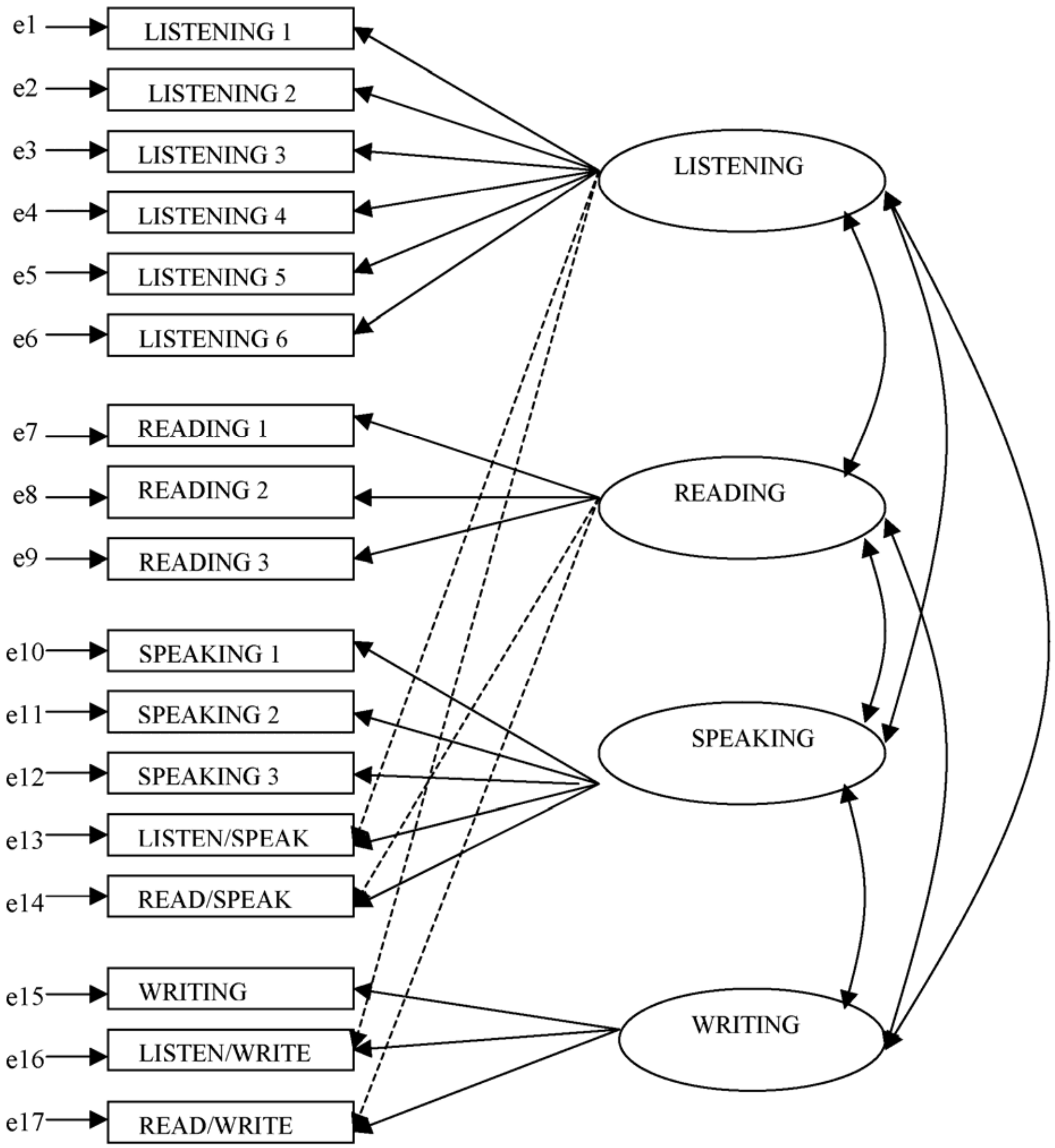

Figure C1. **Alternative Model 1: Four factors—Listening, Reading, Speaking, and Writing.**

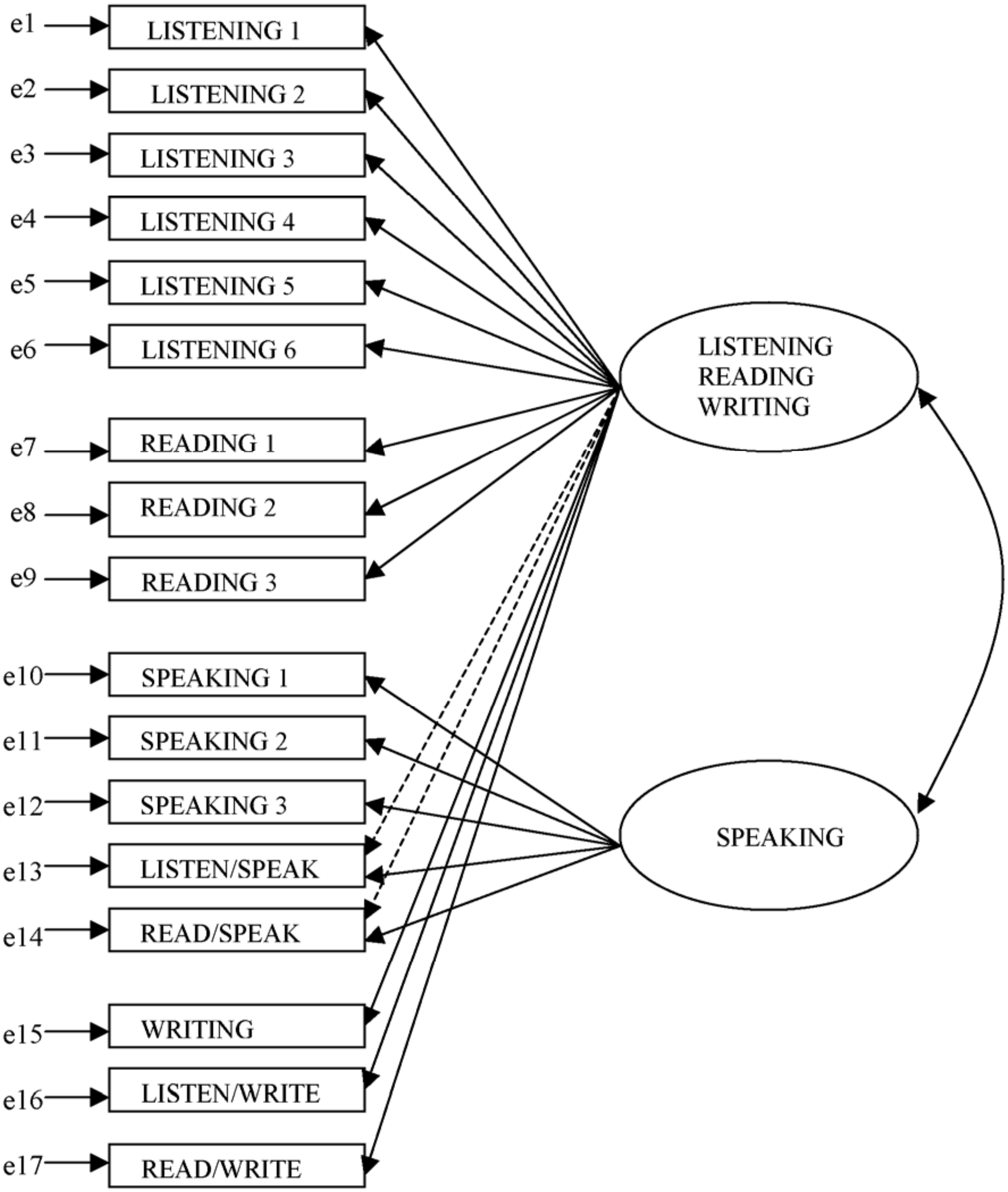

Figure C2. **Alternative Model 3: Two factors—Speaking vs. Listening, Reading, and Writing.**

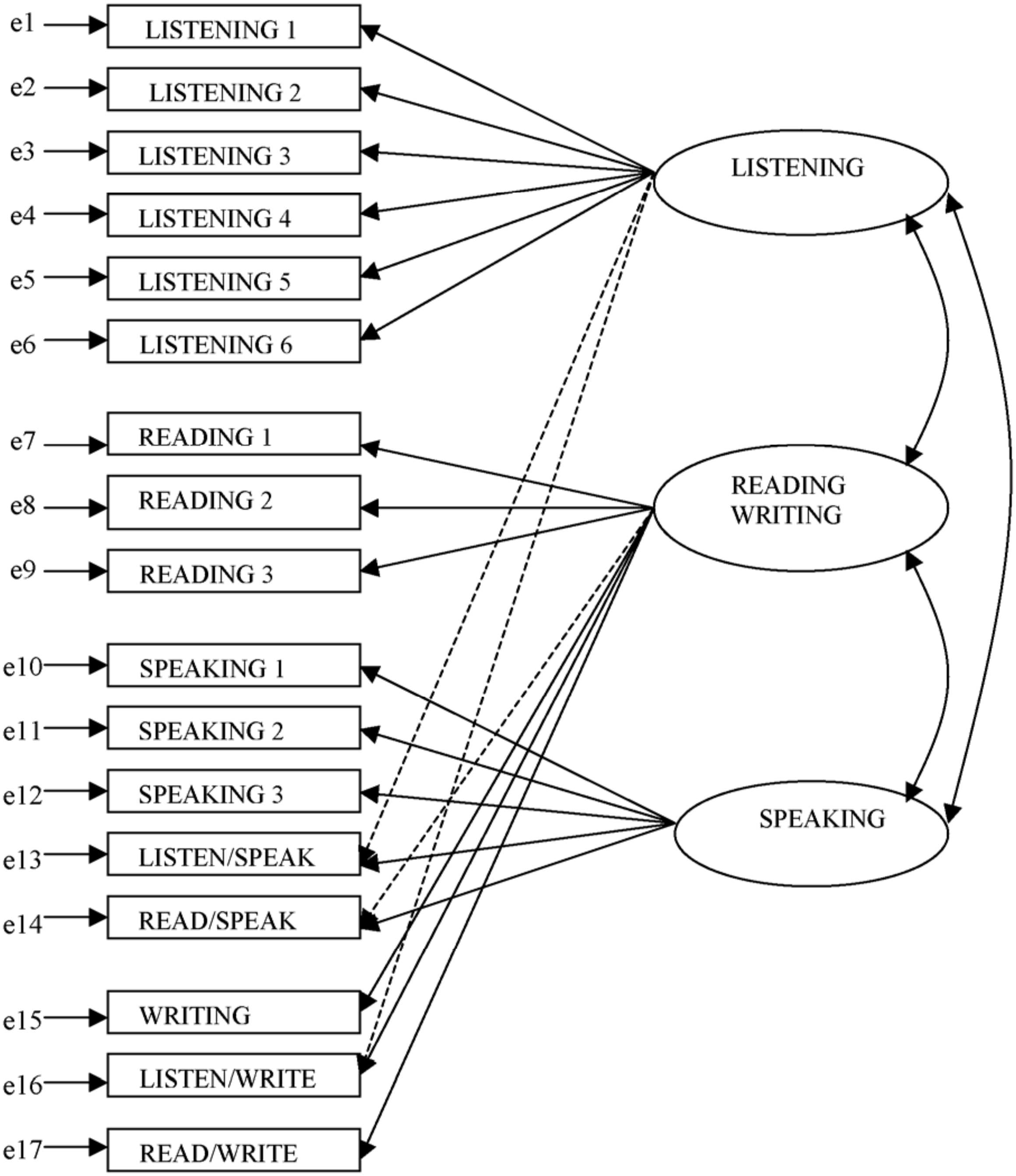

Figure C3. **Alternative Model 4: Three factors—Reading and Writing, Listening, and Speaking.**

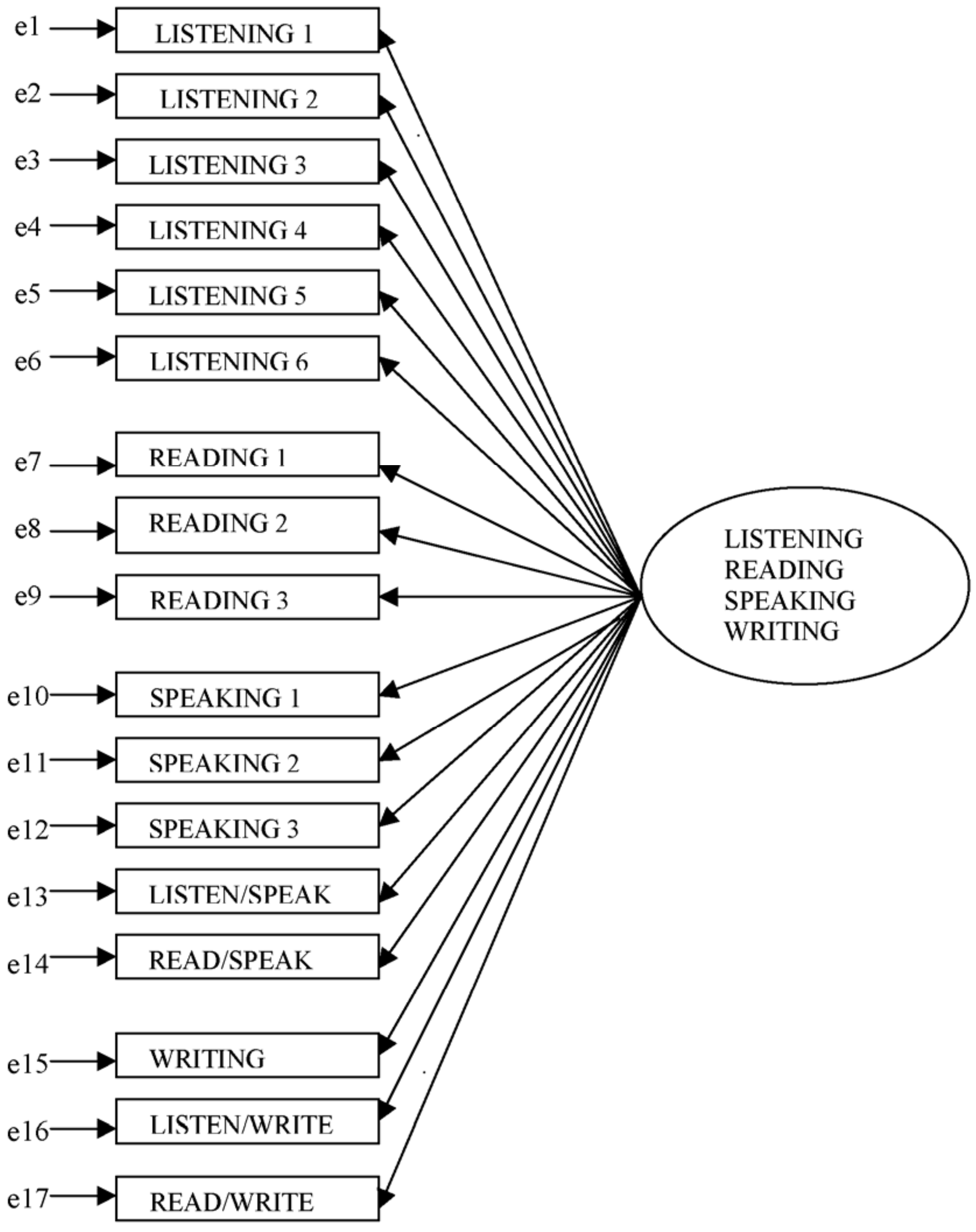

Figure C4. **Original Model 2: One factor—Listening, Reading, Speaking, and Writing.**

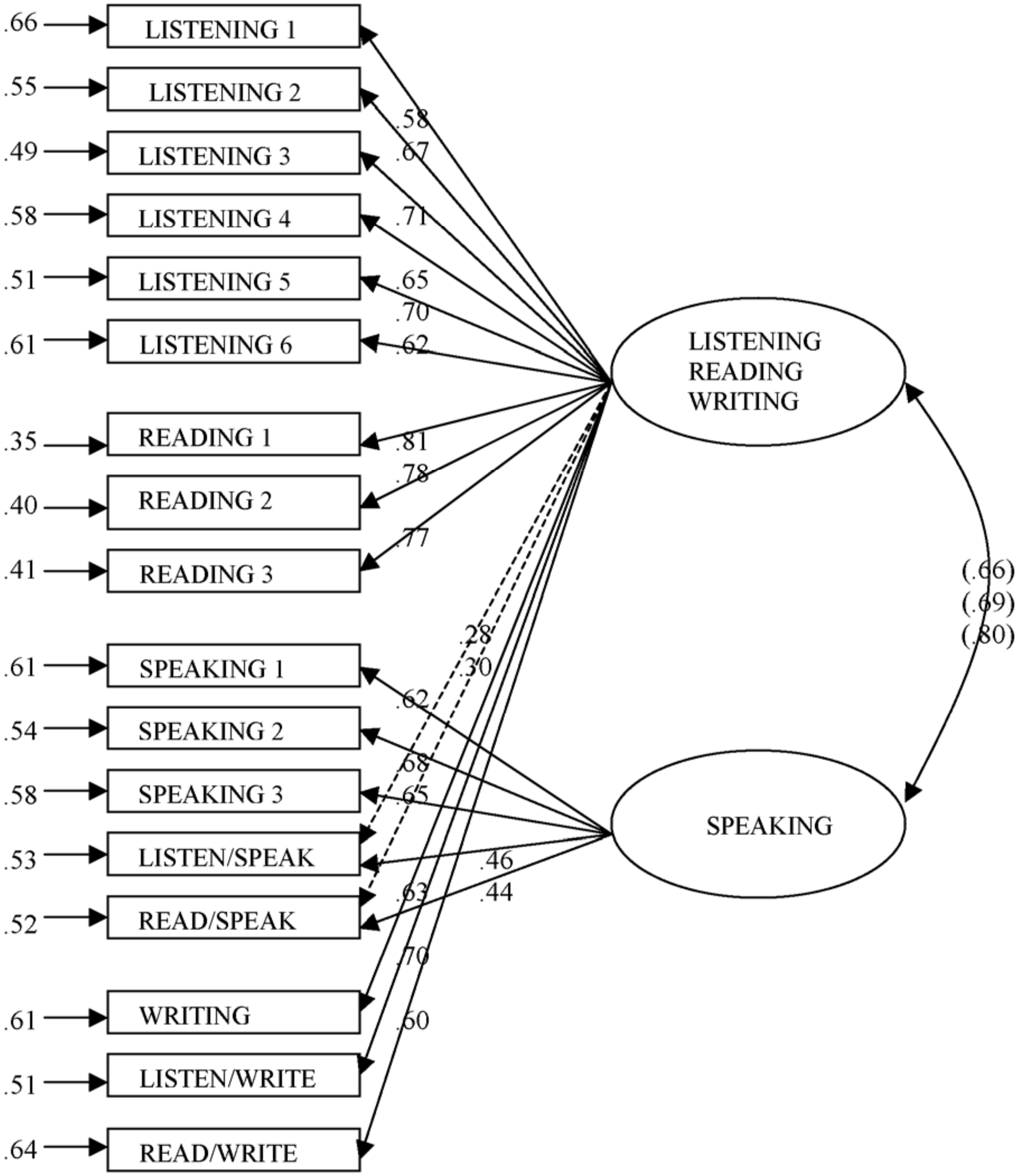

Figure C5. **Alternative Model 3: Two factors—Speaking vs. Listening, Reading, and Writing, with common metric, completely standardized factor loadings and error variances shown.**

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)

Email: toefl@ets.org

Web site: www.ets.org/toefl

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

I.N. 728374