ISSN 1930-9317

TOEFL iBT Research Report

TOEFLiBT-04
March 2008

Factor Structure of the TOEFL Internet-Based Test (iBT): Exploration in a Field Trial Sample

Yasuyo Sawaki

Lawrence Stricker

Andreas Oranje

Factor Structure of the TOEFL® Internet-Based Test (iBT): Exploration in a Field Trial Sample

Yasuyo Sawaki, Lawrence Stricker, and Andreas Oranje
ETS, Princeton, NJ

RR-08-09

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.

Abstract

The present study investigated the factor structure of a field trial sample of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT). An item-level confirmatory factor analysis (CFA) was conducted for a polychoric correlation matrix of items on a test form completed by 2,720 participants in the 2003–2004 TOEFL iBT Field Study. CFA-based multitrait-multimethod (MTMM) analyses for the Reading and Listening sections showed that the language abilities assessed in each section were essentially unidimensional, while the factor structure of the entire test was best represented by a higher-order factor model with a general factor (English as a second language/English as a foreign language ability) and four group factors for reading, listening, speaking, and writing. The integrated Speaking and Writing tasks, which require language processing in multiple modalities, well defined the target modalities (speaking and writing). These results broadly support the current reporting of four scores corresponding to the modalities and a total score, as well as the test design where the integrated tasks contribute only to the scores for the target modalities.

Key words: Construct validity, confirmatory factor analysis, integrated task, score reporting

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

The authors’ special thanks go to Fred Davidson, Neil Dorans, and Richard M. Luecht for their careful review of an earlier version of this research report.

Table of Contents

List of Tables

List of Figures

Introduction

The Test of English as a Foreign Language™ (TOEFL®) is a battery of academic English language ability measures designed primarily for admission of nonnative speakers of English to higher education institutions in North America. The introduction of the TOEFL Internet-based test (iBT) in late 2005 signifies one of the major changes to the test design. In addition to the transition from a computer-based to Internet-based test delivery system, the new test is based on design principles that are drastically different from those of the previous versions. The primary goal of this change is to better align the test design to the variety of language use tasks that examinees are expected to encounter in everyday academic life. Toward this end, a mandatory speaking section has been added to the test, along with integrated tasks that require students to process language in more than one modality (e.g., read a text, listen to a lecture on the same topic, and then write a response on what has been read and heard). The written and spoken texts used in the Reading and Listening sections are longer, and note-taking is allowed throughout the test.

This major transformation of the TOEFL test requires building a validity argument for this new test by gathering various types of empirical evidence. One crucial aspect of this construct validation process is investigating the internal structure of the test to ensure that the relationships among test items and sections correspond with the construct definition, so that test scores can be interpreted appropriately (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Bachman, 2005). Factor analysis can be an important tool for addressing this issue. Some previous researchers have investigated the factor structure of the paper-based TOEFL test (Hale et al., 1988; Hale, Rock, & Jirele, 1989; Manning, 1987; Swinton & Powers, 1980), while the most recent factor analysis study by Stricker, Rock, and Lee (2005) analyzed a prototype of the TOEFL iBT. All of these previous studies have suggested that the TOEFL test taps several correlated, psychometrically distinct traits, though the studies disagree about the number and the makeup of these factors. The design of the TOEFL iBT differs from the versions of the test examined in these previous studies—even from the prototype of the TOEFL iBT studied by Stricker et al. (2005). Because the factor structure of the TOEFL test found in the previous studies may not necessarily hold for this new test design, another investigation of the factor structure of the TOEFL test is required.

The present study investigates the factor structure of the TOEFL iBT with three particular goals in mind. The first goal is to investigate the factor structure of the entire test. Previous factor analyses of English ability measures in multiple language modalities have reached a general consensus that language ability is multicomponential. Thus, it is of theoretical interest to see whether this position is supported for the new test as well. In addition, analysis of the factor structure of the entire test would shed light on an aspect of the TOEFL iBT score reporting policy. In the new test, four scores corresponding to the four sections—Reading, Listening, Speaking, and Writing—are reported, along with a composite score: the total TOEFL iBT score. The policy of reporting the multiple scores corresponding to the four sections and a total score would be supported if the factor structure of the entire test suggests presence of four unidimensional traits corresponding to the four sections and another single dimension that underlies all the sections (i.e., English as a second language/English as a foreign language, or ESL/EFL, ability for academic purposes).

The second goal is to conduct an in-depth analysis of the relationships among the items in the Reading and Listening sections. In the TOEFL iBT, these two sections assess broader skills than their counterparts in the previous versions of the TOEFL. For example, some new item types in the Reading section require examinees to synthesize and organize information presented in a text in order to complete a summary or a schematic table. The Listening section also has similar new items that require examinees to connect information in different parts of the text to complete such tables. These items are specifically designed to tap into skills other than those assessed with more conventional item types, and thus it is of interest to see how these items are related to other item types in these sections. A related question about the Reading and Listening sections concerns the interrelationships among the constructs set forth in the test specifications. The TOEFL iBT test specification identifies three types of academic reading and listening abilities assessed in the Reading and Listening sections, respectively. Although these abilities are psychologically distinct skills and processes, they may not necessarily be psychometrically distinct from one another (Henning, 1992). Thus, it is necessary to empirically investigate the relationships among the different types of language abilities assessed in these two sections, which were constructed based on the TOEFL iBT test specifications.

The third goal is to investigate the relationships between the newly-introduced integrated tasks in the Speaking and Writing sections that require language processing in multiple modalities and the traits assessed in the four sections. The TOEFL iBT includes integrated tasks that combine either two language modalities (Listening/Speaking tasks) or three modalities (Reading/Listening/Speaking and Reading/Listening/Writing tasks). Scores on these tasks contribute only to the sections in which these items are included. For example, Listening/Speaking tasks are part of the Speaking section, and the scores for these items contribute to the Speaking score but not to the Listening score. A legitimate question is whether performance on an integrated speaking or writing task reflects speaking or writing ability, respectively, rather than reading or listening ability. To address this question, the relationships of the integrated tasks to the section scores must be investigated.

Review of Literature

The question of whether language ability is unitary or divisible into components has been of interest to applied linguists for more than 30 years. This issue gained great attention when Oller (1976) proposed the unitary trait hypothesis. Oller proposed the existence of an internalized grammar, or *expectancy grammar,* which allows efficient, online processing of information and creative use of the language. Moreover, because of the similarity in performance of language learners on ostensibly different measures of language ability, he further hypothesized that language ability can be accounted for by a single trait. Strong support for Oller's claim was obtained in a variety of studies conducted by Oller himself as well as his associates, based on principal component analyses of a variety of English language assessments in multiple modalities (e.g., Oller, 1976; Oller & Hinofotis, 1980).

However, Oller's hypothesis was called into question by other researchers (e.g., Carroll, 1983; Farhady, 1983) primarily because of the methodological flaws in the initial studies. Subsequent research in the 1980s and later employed more powerful factor analytic approaches, such as confirmatory factor analysis (e.g., Bachman & Palmer, 1981, 1982; Kunnan, 1995) and exploratory factor analysis with the Schmid-Leiman procedure (1957), which extracts hierarchical factor structures in exploratory factor analysis (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Carroll, 1983). These studies disconfirmed at least an extreme version of the unitary trait hypothesis: only one general factor sufficiently accounts for all of the common variance in language tests.

The current consensus in the field of language testing is that second language ability is multicomponential, with a general factor as well as smaller group factors (Carroll, 1983; Oller

1983). Despite this general consensus, previous findings vary in terms of the exact factor structures that were identified. Some studies found correlated first-order factors, which are often called group factors (e.g., Bachman & Palmer, 1981; Kunnan, 1995), while others found group factors as well as a higher-order general factor (e.g., Bachman & Palmer, 1982; Sasaki, 1996; Shin, 2005). As pointed out by Sasaki (1996), the use of different analysis methods and the characteristics of the populations involved in these studies presumably have contributed to these divergent findings.

Parallel to this line of research in the field of language assessment in general, previous studies of the structure of the TOEFL have also supported the multicomponential nature of language ability. For example, Swinton and Powers (1980), Manning (1987), and two studies conducted by Hale and his associates (Hale et al., 1988, 1989) examined the factor structure of the paper-based TOEFL test, which consisted of three sections: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. Swinton and Powers (1980) and Manning (1987) employed an exploratory factor analysis, while the two studies conducted by Hale and his associates used confirmatory factor analysis. These studies differed in the type of data analyzed as well. Swinton and Powers (1980) analyzed item-level data, while all the other three studies conducted their factor analyses on item parcels (i.e., groups of items).

Despite these methodological differences, all four studies found a distinct Listening Comprehension factor. However, the studies differed in the number and makeup of the other factors that they identified. Swinton and Powers (1980) found three correlated factors—a Listening factor and two other factors defined by different combinations of Structure, Written Expression, Vocabulary, and Reading Comprehension across different language groups—for all the seven language groups studied. In contrast, Hale et al. (1988) found two correlated factors, one for Listening Comprehension and the other for Structure, Written Expression, Vocabulary, and Reading Comprehension. Manning's (1987) results were similar to those of Hale et al.'s study, in that they identified a distinct Listening Comprehension factor that was moderately correlated with the other, more general factor primarily defined by the Structure and Written Expression as well as the Reading Comprehension and Vocabulary sections. These divergent findings were further explored by Hale and his associates (1989). In this study, the data from domestic and overseas test centers from a 1976 administration studied by Swinton and Powers as well as those from a 1984 administration were re-analyzed (Hale et al., 1988, included only the domestic portion of the 1984 data). A confirmatory factor analysis of item parcels found that the same two-correlated-factor solution observed by Hale et al. (1988) was generalizable across domestic and overseas test takers, five native language groups, and the 1976 and 1984 test forms.

A recent multiple-group confirmatory factor analysis by Stricker et al. (2005) studied the factor structure of a prototype of the TOEFL iBT, called LanguEdge Courseware, for three language groups. Similar to the TOEFL iBT structure, this prototype consisted of four sections (Reading, Listening, Speaking, and Writing). Item parcels for the multiple-choice items in the Reading and Listening sections and the holistic ratings obtained for individual Speaking and Writing items were analyzed. A correlated two-factor model—one factor for Speaking and the other factor for a fusion of Reading, Listening, and Writing—was identified for all three language groups. A simultaneous analysis of this model for the three groups also suggested invariance of factor loadings and error variances, but differences in the correlations between the two factors across the three language groups. Stricker et al. (2005) concluded that the relative distinctness of speaking observed in their study may reflect the effects of instruction. Because speaking was not a mandatory component of the TOEFL test in the previous versions, speaking may have been de-emphasized, resulting in the emergence of the distinct speaking factor. Thus, the factor structure of the TOEFL iBT may look quite different in a future study, if the introduction of the new test leads to more emphasis on speaking instruction.

Given the similarity of the TOEFL iBT design to that of LanguEdge, the results of the present study may be somewhat similar to those of Stricker et al. (2005). However, the design of TOEFL iBT is not identical to that of LanguEdge, as described below. In addition, the sample size of the Stricker et al. (2005) study was rather modest (the entire sample consisted of data from 439 examinees in total, where the sample sizes for the language groups varied from 100 to 225). For these reasons, it is possible that the results of this study may depart from those of Stricker et al.

Method

Data

The data analyzed in the present study were scored item responses in the Reading, Listening, Speaking, and Writing sections of a TOEFL iBT test form (Form A) administered as part of the TOEFL iBT Field Test conducted in November 2003 through February 2004. The paid participants were recruited from 31 countries in North America, Latin America, Africa, Asia, and Europe that accounted for about 80% of the 2001–2002 TOEFL testing volume (Chapelle, Enright, & Jamieson, 2008). During the field test session, all participants were required to complete the new TOEFL iBT test form (Form A) and a TOEFL computer-based test (CBT) form. In addition, a fraction of the sample completed an additional form of the TOEFL iBT (Form B). The iBT and CBT test forms were administered to the participants in a counter-balanced fashion. Each participant filled out an online questionnaire of demographic, self-assessment, and post-test questions as well.

In total, 2,720 usable responses were available from the TOEFL iBT Form A. Based on the participants' reported native country information from the 2002-2003 TOEFL candidates, this sample was reasonably representative of the operational TOEFL population in terms of reported native countries of origin (Chapelle et al., 2008). The five largest groups were from India (14.8%), China (13.9%), South Korea (10.1%), Japan (7.5%) and Taiwan (4.6%). Approximately 9% of the examinees did not report their native countries. Of all the 2,720 participants, 672 (24.7%) completed the study materials at domestic (United States and Canada) test centers, and 2,048 (75.3%) did so at overseas test centers.

The participants' TOEFL CBT scores provide information about their English language ability levels. The scaled CBT test scores are summarized in Table 1 along with comparable data for CBT test takers in 2002–2003. The field test sample was approximately half a standard deviation below the CBT test-taking population on all the CBT subscores and the total scores. Results of one-sample *t* tests for the section and total scores showed that all of the means for the field test sample were significantly lower than those for the CBT population (Listening: $t = 28.17$, $p < .05$, $df = 2719$; Reading: $t = 30.51$, $p < .05$, $df = 2719$; Structure & Writing: $t = 17.91$, $p < .05$, $df = 2719$; Total: $t = 28.35$, $p < .05$, $df = 2719$). Moreover, the obtained Cohen's (1988) *d* values indicated that the observed effects were medium for the Listening ($d = .54$), Reading ($d = .59$) and Total ($d = .54$) scores, while that for the Structure and Writing ($d = .34$) score was small. This finding should be kept in mind when considering the generalizability of the results to the TOEFL test-taking population.

In accordance with the design of the operational TOEFL iBT test form, Form A consisted of four sections: Reading, Listening, Speaking, and Writing. A summary of the features of the entire test and the individual test sections is presented in Table 2.

Table 1

Summary Statistics on TOEFL iBT and TOEFL CBT Scores

	TOEFL iBT Field Test sample (N = 2,720)		TOEFL population (N = 577,038)[a]	
Scaled scores	*M*	*SD*	*M*	*SD*
CBT Listening	17.7	6.0	20.9	5.3
CBT S/W	19.5	6.5	21.7	5.0
CBT Reading	18.3	6.0	21.8	4.9
CBT total	184.8	55.5	215.0	46.0
iBT Reading	17.0	7.0	-	-
iBT Listening	17.0	7.0	-	-
iBT Speaking	17.0	7.0	-	-
iBT Writing	16.0	6.7	-	-
iBT total	67.0	24.6	-	-

Note. CBT = Computer-based test; iBT = Internet-based test; S/W = Structure/Writing.

[a] The statistics are based on the performance of the total group of 577,038 examinees tested between July 2002 and June 2003 in computer-based testing test centers (ETS, 2003).

Table 2

Structure of TOEFL iBT Form A

Section	Items	Constructs	Other features	Scores
Reading (39 items)[a]	36 dichotomous items 3 polytomous items	3 purposes of reading	3 sets	45 maximum raw score points Reading section score on a scale of 0–30
Listening (34 items)[b]	32 dichotomous items 2 polytomous scored items	3 listening abilities	Conversation: 2 sets Lectures: 4 sets	35 maximum raw score points Listening section score on a scale of 0–30
Speaking (6 items)	6 items scored on a 5-point rating scale	2 items: Independent Speaking 2 items: Listening/Speaking 2 items: Reading/Listening/Speaking		24 maximum raw score points Speaking section score on a scale of 0–30
Writing (2 items)	2 items scored on a 6-point rating scale	1 item: Independent Writing 1 item: Reading/Listening/Writing		10 maximum raw score points Writing section score on a scale of 0-30
TOEFL total scaled score				0-120

[a] One Reading item that was not scored was excluded from the subsequent analyses. [b] One Listening item that was not scored was excluded from the subsequent analyses.

Structure of the Test

Reading. The Reading section consisted of three item sets, each of which contained 12-14 items associated with a common reading passage of approximately 700 words in length. The examinees were allowed to spend 60 minutes to complete the Reading section. Thirty-six items were dichotomously scored, four-option multiple-choice items, and the remaining three were polytomously scored items. One dichotomously scored item was excluded during an initial item analysis because of its poor performance. After excluding this item, 38 items were included in the subsequent analyses. The raw scores for this section ranged from 0 to 45.

Listening. The Listening section consisted of six item sets, two based on conversations, and four based on lectures on academic topics. Each conversation stimulus was approximately three minutes long and was followed by 5 multiple-choice items, while each lecture stimulus was 3–5 minutes long and was followed by 6 multiple-choice items. In addition to the time required for listening to the prompts, the participants were allowed to spend up to 20 minutes to respond to all Listening items. Thirty-two of the 34 items were scored dichotomously, while the remaining two were worth more than one point each. One dichotomous item identified as misfitting in an IRT analysis was removed from further analyses. Thirty-three items were included in the subsequent analyses after removing this item. The raw scores for this section ranged from 0 to 35.

Speaking. The Speaking section consisted of six tasks. Two were independent tasks, which required examinees to express opinions on familiar topics. The other four were integrated tasks. Two of the four were Listening/Speaking tasks, which required examinees to listen to a short spoken text and then respond to it. The remaining two were Reading/Listening/Speaking tasks, which required examinees to read a short text, listen to a spoken text that pertained to the reading text, and then respond about what they had read and heard. For each task, examinees were given 15–30 seconds to prepare, and 45–60 seconds to respond. Each examinee’s response to each task was scored on a scale of 0–4 by trained raters. Each speaking response of those students who completed Form A only was scored by a single rater, and the score given by the rater was the final score for that task. Each speaking response of a subsample of students who completed both Forms A and B was rated by two raters. When there were discrepancies of more than one point, a chief rater provided the final score for the task. The raw Speaking section score was a sum of the points earned on the six tasks. The raw scores for this section ranged from 0 to 24.

Writing. The Writing section included two tasks, one of which was an independent writing task, and the other of which was an integrated writing task. The independent writing task required examinees to support an opinion on a common topic. The integrated writing task required them to read a text, listen to a lecture that pertained to the topic, and then respond to a specific question on what they had read and heard. For each question, examinees were required to type their answers. The total testing time for the Writing section was 50 minutes, 20 minutes allocated to the independent writing task and 30 minutes to the integrated writing task. Each examinee's response to each task was scored on a scale of 0 to 5 by two trained raters. The final score on the item was the average of the scores of the two raters in half point intervals. If the ratings provided by the two raters differed by more than one point, however, a third rater scored the response for adjudication. The final task score was the average of the three scores if they were adjacent to one other. If not, the final task score was the average of the two most adjacent scores among the three.[1] The raw Writing section score was a sum of the points earned on the two items. The raw scores for this section ranged from 0 to 10.

For each section, the raw section scores were converted to scaled scores. By using a linear transformation method, the raw section scores were brought to a scale of 0–30 with the same scaled mean and standard deviation across the sections. The total score, a simple sum of the four scaled section scores, was on a scale of 0–120. The scaled TOEFL iBT scores for the 2,720 field test participants in the present sample are presented in Table 1.

Form A was representative of the content and format of operational TOEFL iBT test forms, except for two points. First, three out of four lecture item sets included in the Form A Listening section were based on nonscience lectures, while the Listening section in each operational TOEFL iBT test form contains two science and two nonscience lecture item sets. Second, in operational TOEFL iBT examinees may receive additional Reading or Listening item sets. Two important differences in the test administration conditions must be noted as well. Unlike in the field test, where the three Reading item sets and the six Listening item sets were administered consecutively, the item sets in the Reading and Listening sections in the operational TOEFL iBT are administered in separately-timed subparts. Moreover, in the Form A Listening section, the two conversation item sets preceded the four lecture item sets, while in operational TOEFL iBT, the item sets are presented in the order of Conversation-Lecture-Lecture within each of the two subparts.

Grouping Items Within the Reading and Listening Sections

Besides the macro-level characterization of the test design discussed above, the participants' responses to the multiple-choice items in the Reading and Listening sections can be investigated at a more detailed level by grouping items on the basis of some classification scheme. Several classification schemes are provided. One empirical approach is to conduct an exploratory factor analysis and categorize items according to the factors on which they load. Another approach is to cluster items by item sets (i.e., a group of items based on the same reading or listening passage). This method was used in the factor analysis of the LanguEdge Courseware by Stricker et al. (2005). This approach is designed to alleviate often encountered inherent instabilities in factor analyses of item-level data and may eliminate artifacts due to local dependence. Other approaches are based on content considerations. Items can be classified according to the definitions of the abilities assessed in the Reading and Listening sections based on the test specifications. Items can also be classified on the basis of the scheme developed for cognitive diagnosis (Nissan & Sawaki, 2005; Sawaki & Lee, 2006; von Davier, 2005; Zhang, DiBello, Puhan, Henson, & Templin, 2006). This scheme was based on a task analysis of the Reading and Listening items by content experts. This scheme identified four skills for both the Reading and Listening sections.

In the present study, the test specification classification was employed. The test specifications for the TOEFL iBT Reading and Listening sections were derived from the construct definitions set forth in the framework papers on the development of these sections (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Enright et al., 2000).[2] Originally, the Reading framework paper identified four purposes of reading, or four different aspects of reading subsumed under the broad construct of reading comprehension in academic settings. This four-part classification was reorganized into three in the test specifications, which further defined item types designed to assess abilities associated with each type of reading as well as task parameters that guided the test's item development. Below are the three purposes of academic reading and their associated item types:

- Basic Comprehension—Vocabulary, Reference, Sentence Simplification, Factual Information, and Negative Fact
- Reading to Learn—Prose Summary, Classifying/Categorizing/Organizing Information

- Inferencing—Inference, Rhetorical Purpose, Insert Text

Similarly, the Listening framework paper (Bejar et al., 2000) described early conceptualizations of the types of academic listening as well as task characteristics for the design of the TOEFL iBT Listening section. The current form of the test specifications for the Listening section identifies three aspects of academic listening ability: Basic Understanding, Pragmatic Understanding, and Connecting Information.

Form A included 26 Basic Comprehension items (26 points), nine Inferencing items (9 points), and three Reading to Learn items (9 points) in the Reading section.[3] The Listening section included 17 Basic Understanding items (17 points), six Pragmatic Understanding items (6 points), and ten Connecting Information items (12 points).[4]

Analyses

After a series of preliminary analyses, various CFA models were tested in the main analyses to address the three research questions laid out above. The first series of CFAs focused on in-depth analysis of the factor structure of the sections. Separate analyses were conducted for the Reading and Listening sections, while the Speaking and Writing sections were analyzed together because of the small number of items in each section. Then, the CFA models for the Reading, Listening, and Speaking and Writing sections obtained above were combined in order to test the factor structure of the entire TOEFL iBT and to investigate the relationships of the integrated Speaking and Writing items with the four TOEFL sections.

The present study employed an item-level confirmatory factor analysis. A CFA with ordinal categorical data is appropriate for factor analysis of item responses where each item is scored dichotomously or polytomously. In this case a polychoric correlation matrix of the item response data is analyzed, unlike a conventional CFA of continuous variables (e.g., test section scores) that analyzes a variance-covariance matrix. This item-level factor analysis has been applied to analysis of language assessment data by a few previous researchers (e.g., Carr, 2003; Davidson 1988; Swinton & Powers, 1980). This approach was used to address two of this study’s goals: to conduct a fine-grained analysis of the relationships among individual items in the Reading and Listening sections and to investigate the relationships of the integrated items to the other parts of the test.

The polychoric correlation matrix for the item-level data was calculated using PRELIS 2.54 (Jöreskog & Sorbom, 2003a), and LISREL 8.54 (Jöreskog & Sorbom, 2003b) was employed for the series of CFAs.

Data cleaning and preliminary analyses. At the outset descriptive statistics for the items and section scores, and total scores (raw and scaled scores) were examined. One important goal of these item analyses was to identify items with extremely high or low item difficulty values, which could be problematic in the calculation of a polychoric correlation matrix (McLeod, Swygert, & Thissen, 2001). After these preliminary analyses, a polychoric correlation matrix for all the Form A items was obtained.

CFA within the Reading and Listening sections. The input data for all the CFAs consisted of the polychoric correlation matrix for the Form A items. In the separate analyses of the Reading and Listening sections, a series of CFA models for multitrait-multimethod (MTMM) analyses (Jöreskog, 1974; Marsh, 1988, Marsh & Grayson, 1995; Widaman, 1985) were tested. The CFA approach to MTMM analysis is the most commonly used alternative to Campbell and Fiske's (1959) original MTMM analysis based on an observed correlation matrix. This in-depth investigation of the trait and method factor structure within the Reading and Listening sections was motivated by the finding in the preliminary EFAs for these sections that there were non-negligible method effects associated with item sets and/or item locations in both sections (hereafter, item set effects). The main goal of the MTMM analysis was not to maximize model fit by taking into account the score variability due to the item set effects. Rather, the aim was to evaluate the size of the item set effects and investigate the feasibility of testing more parsimonious models that specify only trait factors for the entire test. This approach avoids adoption of an overly complex CFA model that may not be replicable in another sample.

Widaman (1985) and Marsh (1988) proposed a taxonomy of CFA models to be tested in an MTMM analysis. Although not all of the models in the taxonomy were tested in this study, the four key models that Marsh and Grayson (1995) recommended were investigated. The proposed models were evaluated in terms of the appropriateness of the solutions, theoretical interpretability of the results, and goodness of model fit to the data. For the evaluation of model fit, both statistical tests (model chi-square and chi-square differences between alternative models being compared) as well as indices of model fit were considered.

For estimating model parameters, maximum likelihood estimation was used. Because the input data are categorical and non-normal, an asymptotic covariance matrix for the same sample obtained in PRELIS was read in LISREL, with the polychoric correlation matrix for computation of the fit indices adjusted for the non-normality of the data. This allowed calculation of the Satorra-Bentler scaled chi-square statistic (Satorra, 1990). The goodness-of-fit criteria described below, largely based on Hoyle and Panter's (1995) suggestions, were used in this study:

- *The ratio of Satorra-Bentler model chi-square to model degrees of freedom ($\chi 2_{S\text{-}B}/df$).* Although there is no clear-cut rule about a cutoff point for this statistic, Kline (1998) mentions 3.0 or below as a suggestion of good model fit.

- *Goodness of fit index (GFI).* An absolute model-fit index, which is analogous to a model R^2 in multiple regression analysis. A GFI of .90 or above indicates an adequate model fit.

- *Non-normed fit index (NNFI).* An incremental fit index, NNFI is an extension of the Tucker-Lewis index (TLI). An NNFI assesses whether a particular CFA model is an improvement over a model that specifies no latent factor, taking into account the model complexity (Raykov & Marcoulides, 2000). An NNFI of .90 or above indicates an adequate model fit.

- *Comparative fit index (CFI).* An incremental fit index, which assesses overall improvement of a proposed model over an independence model where the observed variables are uncorrelated. A CFI of .90 or above indicates an adequate model fit.

Besides the indices above, two more criteria below were also taken into account:

- *Root mean square error of approximation (RMSEA).* An RMSEA evaluates the extent to which the model approximates the data, taking into account the model complexity. A RMSEA of .05 or below is considered as an indication of close fit, and a value of .08 or below as an indication of adequate fit (Browne & Cudeck, 1993).

- *Expected cross-validation index (ECVI).* An ECVI indicates the extent to which the model is replicated with a different sample from the same population. The lower the value, the better the replication of the result in another sample.

In addition, when considering the number of distinct factors present in the data, the magnitudes of interfactor correlations were evaluated. When a correlation between two factors is extremely high, the two factors cannot be considered to be distinct from each other (Bagozzi & Yi, 1992). A correlation of .90 was chosen as the rule of thumb for the sake of consistency with the criterion used by Stricker et al.'s (2005) CFA study of the LanguEdge Courseware data.

In addition to the evaluation of overall goodness of fit of individual CFA models, relative goodness of fit of competing models were compared by a sequential building of the CFA models for the entire test. For comparison of two nested models, a chi-square difference test was conducted with Satorra and Bentler's (1999) adjustment procedure for the use of the Satorra-Bentler scaled model chi-square statistic. Chi-square difference test results for model comparisons were always evaluated in conjunction with the subjective goodness-of-fit criteria listed above (i.e., $\chi^2_{S\text{-}B}/df$, GFI, NNFI, CFI, RMSEA, and ECVI).

CFAs for the Speaking and Writing sections. The data from the Speaking and Writing sections were combined to conduct another series of CFA analyses for the two sections together. The models tested for the Speaking and Writing sections did not involve any factors associated with the test method, such as the item set effects considered for the Reading and Listening sections. Rather, the purpose of the analysis was to identify the trait factor structure that best represented the relationships among the Speaking and Writing items. The model testing procedure used for the Reading and Listening section analyses was followed.

CFAs for the entire test. The CFA models developed separately for the Reading, Listening, and Speaking and Writing sections above were combined to build CFA models for the entire test. The primary focus of this analysis was to explore the trait factor structure of the entire test. At this stage, relative goodness of fit of a series of nested CFA models representing different trait factor structures of the test were compared, following Rindskopf and Rose's (1988) procedure. The procedures for the section-level CFAs were followed.

Results

Preliminary Analyses

Item analyses and descriptive statistics. For both the Reading and Listening sections, there was a large variation in the obtained *p*-values for the dichotomously scored items, ranging from .40 to .94 for Reading, and .31 to .88 for Listening. However, none of the items were associated with extremely low or high *p*-values. An inspection of the *p*-values for different item types suggested that none of the TOEFL framework item types was either consistently more difficult or easier than another.

The remaining two TOEFL iBT sections—Speaking and Writing—were based on polytomously scored items based on holistic rating scales. Across all the six Speaking items, the means were fairly homogeneous, ranging from 2.12 to 2.52 (standard deviations, 0.96 to 1.21, respectively). The means for the Independent Writing and the Reading/Listening/Writing tasks were 2.14 and 2.98 (standard deviations, 1.43 and 1.27, respectively).

Finally, polychoric correlation coefficients based on the responses of the 2,720 Field study participants were obtained for the item-level data for all the four sections. All the 79 items in the four sections were retained in the polychoric correlation matrix as well as the subsequent analyses.

Confirmatory Factor Analyses of the Sections

Reading section. Three trait and three item set factors were considered throughout the MTMM analyses of the Reading section data. The trait factors represented the three purposes of reading identified in the test specification for the Reading section: Basic Comprehension, Inferencing, and Reading to Learn.[5] Moreover, the trait factors were specified as being correlated with one another. Some degree of correlation is expected among different measures of ability to read in English. The three item set factors corresponded to the three item sets. Whether the item set factors were specified as correlated or uncorrelated depended on the particular models tested, as discussed below. Furthermore, the trait factors were specified as uncorrelated with the item set factors.

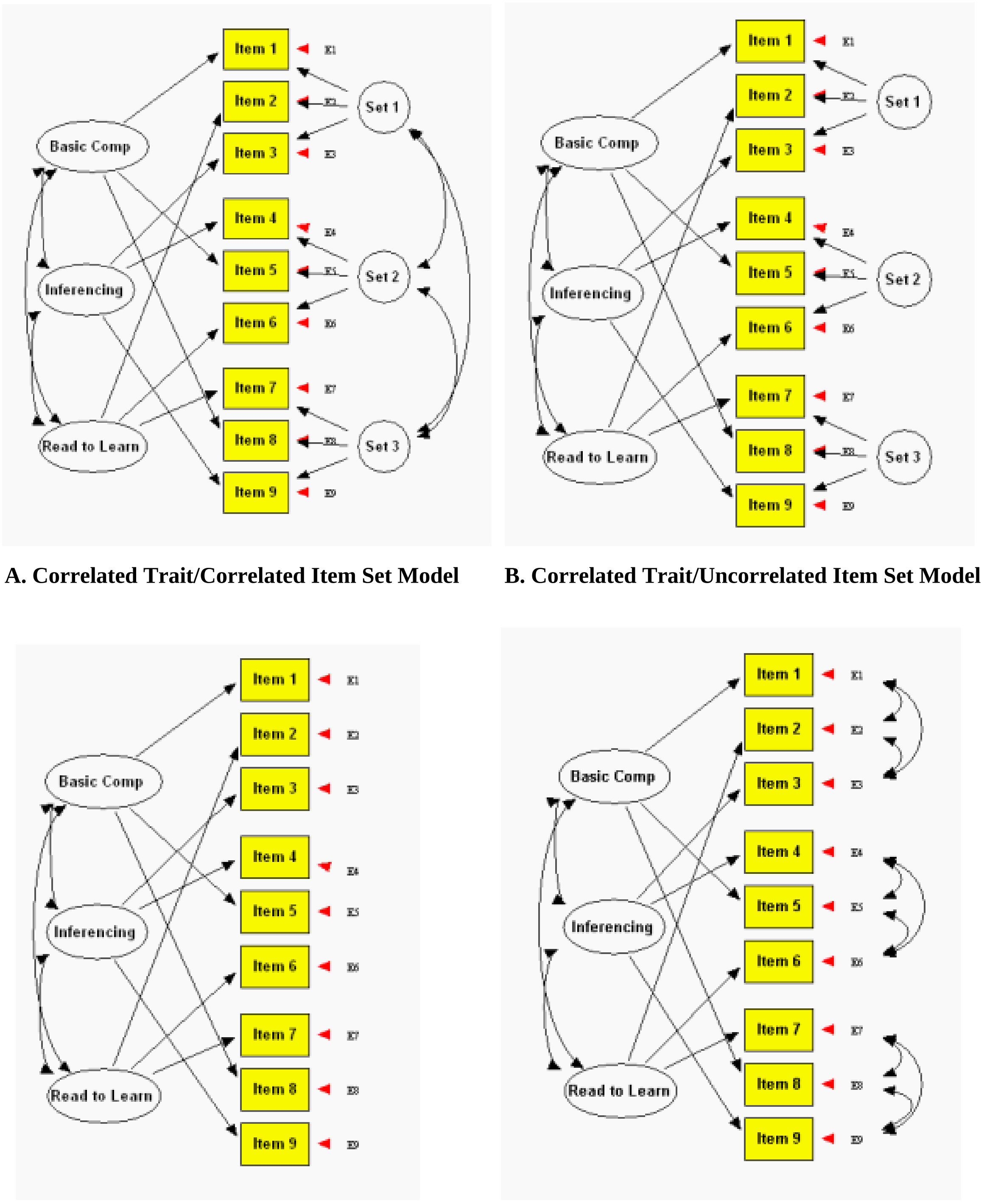
Initially a series of MTMM models involving the three traits and the three item set factors (see Figure 1 for sample schematics) were tested along the lines suggested by Marsh and Grayson (1995):

1. Correlated trait/correlated item set model (Panel A): All of the three trait factors and the three item set factors were modeled. Not only the trait factors but also the three item set factors were modeled as correlated among themselves. This model is considered as a full model, against which relative goodness of fit of the other models was considered.

2. Correlated trait/uncorrelated item set model (Panel B): This model was nested within the correlated trait/correlated item set model above and was obtained by fixing the correlations among the item set factors in the full model to zero. A comparison of this model against the full model would show the extent to which the item set factors are correlated.
3. Correlated trait model (Panel C): This model was nested within the two models above as well as the correlated trait/correlated uniqueness model below. This model was obtained by completely trimming the item set factor structure of the full model. A comparison of this model against the other three models would indicate the degree to which item set effects are present in the data.
4. Correlated trait/correlated uniqueness model (Panel D): This model had the identical correlated trait factor structure as the three other models, but no item set factors were specified. Instead, covariances among the residuals of the items associated with the same item set were estimated. The residual correlations reflect a combination of the effect of the item set and error.

The last model, correlated trait/correlated uniqueness model, involves estimation of a considerably larger number of model parameters compared to the other three models. Moreover, modeling correlated residuals in the model leads to difficulty in replicating the results in another sample and interpreting the results. Thus, it is not a preferred choice when considering model parsimony. This model was tested, however, because previous studies showed that this model is more likely to result in proper solutions (i.e., a solution is proper if the model is identified and associated with estimated parameters within their permissible ranges) compared to other conventional MTMM models such as the correlated trait/correlated item set and correlated trait/uncorrelated item set models above (Marsh & Bailey, 1991; Marsh & Grayson, 1995). Thus, the parameter estimates for this model were used primarily as guidelines against which the appropriateness of the other three solutions were evaluated rather than considering the model as a likely candidate for a final model.

The solutions based on these four models were examined against the three criteria suggested by Marsh (1988) and Marsh and Grayson (1995): (a) the extent to which the solution was proper, (b) substantive interpretability, and (c) goodness of fit of the model.



A. Correlated Trait/Correlated Item Set Model

B. Correlated Trait/Uncorrelated Item Set Model

C. Correlated Trait Model

D. Correlated Trait/Correlated Uniqueness Model

***Figure 1*. Schematic representation of the four initial MTMM models tested for the Reading section.**

Note. Basic Comp = Basic Comprehension; Read to Learn = Reading to Learn.

All four models satisfied the first criterion listed above. However, the solution for the correlated trait/correlated item set model was not substantively interpretable: a majority of the items had low positive trait factor loadings and large positive method factor loadings. Thus, the correlated trait/correlated method model was not considered further. The solutions for the remaining three models were proper. Moreover, across all three models, the patterns observed in the factor loadings were similar: moderate positive trait factor loadings, low and positive method factor loadings, and high inter-trait-factor correlations. These patterns observed in the trait factor loadings and correlations were substantively reasonable, considering the expected psychometrically unidimensional nature of the reading ability assessed in the TOEFL Reading section (Hale et al., 1988, 1989; Schedl, Gordon, & Tang, 1996; Swinton & Powers, 1980). The goodness of fit of these models was satisfactory as well. Although these models were all possibly amenable to further interpretation, satisfying the second and third criteria, one problem common across these three models was the extremely high interfactor correlations, equal to or above .96, which did not meet the predetermined criterion of an interfactor correlation of .90 or below to claim that two given constructs are distinct from each another.

Consequently, another series of models analogous to the correlated trait/correlated item set, correlated trait/uncorrelated item set, correlated trait, and correlated trait/correlated uniqueness models was obtained by fixing all the inter-trait-factor correlations in the initial four models to 1.0. This is equivalent to specifying the three trait factors as being psychometrically indistinguishable (i.e., all the items within the Reading section load onto a single trait factor).

The new set of models with a single trait factor corresponding to the correlated trait/correlated item set, correlated trait/uncorrelated item set, correlated trait, and correlated trait/correlated uniqueness models were named, respectively, as single trait/correlated item set, single trait/uncorrelated item set, single trait, and single trait/correlated uniqueness models. These four models were all identified and converged with model parameter estimates within their permissible ranges. However, the model parameters for the single trait/correlated item set model were substantively uninterpretable, with primarily moderate negative trait factor loadings and moderate positive item set factor loadings, the pattern of which was drastically different from those observed for the other three models. Thus, the single trait/correlated item set model was ruled out from further consideration. Because the good overall fit of the remaining three models supports the psychometric unidimensionality of the traits assessed in the Reading section, further model comparisons were conducted based on the single trait/uncorrelated item set, single trait and single trait/correlated uniqueness models (see Tables 3–7 for the model parameters for these three models).

Next, relative goodness of fit of the single trait/uncorrelated item set, single trait, and single trait/correlated uniqueness models were compared. First, the goodness of fit of these models was examined based on the goodness-of-fit criteria. As can be seen in Table 8, the single trait/correlated uniqueness model indicated a good fit to the data, satisfying the predetermined criteria for GFI, NNFI, and CFI and having the lowest ECVI value. The single trait/uncorrelated item set model showed an acceptable fit as well, satisfying all criteria, with the exception of the GFI value. Being a more restricted version of both the single trait/uncorrelated item set model and the single trait/correlated uniqueness model, the fit of the single trait model was relatively worse than that of the other two. Nevertheless, the fit of the single trait model appears to be marginally acceptable with the satisfactory NNFI, CFI, and RMSEA values despite the $\chi^2_{\text{S-B}}$/df (3.12) and GFI (.84).

Second, formal comparisons of the three models were conducted. The single trait/uncorrelated item set and single trait models were compared to evaluate the extent to which item set effects are present in the Reading section. A chi-square difference test based on the Satorra-Bentler scaled model chi-square values suggested that the fit of these two models was significantly different ($p < .01$, $\chi^2_{\text{S-B difference}} = 2{,}074.00$; $df = 38$). Despite that, except for the noticeable decrease of GFI by .05, the changes in the NNFI, CFI, and RMSEA values were minimal. This result suggests that, practically speaking, the single trait model, which specifies only the trait factor structure, fits as well as the single trait/uncorrelated item set model.

Next, the single trait and single trait/correlated uniqueness models were compared. This is similar to the comparison of the single trait model with the single trait/uncorrelated item set model, but the difference is that the single trait/correlated uniqueness model does not assume unidimensionality of the method effects (Marsh & Grayson, 1995). A chi-square difference test showed that the fit of the single trait/correlated uniqueness model was significantly better than that of the single trait model ($p < .01$, $\chi^2_{\text{S-B difference}} = 1{,}386.09$; $df = 224$). The subjective criteria of goodness of fit of the models indicated the better fit of the correlated trait/correlated uniqueness model as well.

Table 3

Standardized Parameter Estimates for the Single Trait/Uncorrelated Item Set Model (Reading)

Item	Trait	Item set			Error[a]	SMR
	Reading[a]	Set 1	Set 2	Set 3		
R1	.53	.16[a]			.69	.31
R2	.56	-.13[a]			.67	.33
R3	.67	.19[a]			.52	.48
R4	.67	.00			.55	.45
R5	.47	-.06			.78	.22
R6	.59	.18[a]			.62	.38
R7	.47	.07			.77	.23
R9	.72	-.23[a]			.43	.57
R10	.31	.06			.90	.10
R11	.52	-.05			.73	.27
R12	.73	-.28[a]			.39	.61
R13	.56		.20[a]		.65	.35
R14	.66		.38[a]		.42	.58
R15	.38		.17[a]		.83	.17
R16	.65		-.10[a]		.56	.44
R17	.62		-.23[a]		.56	.44
R18	.55		-.26[a]		.64	.36
R19	.76		-.14[a]		.40	.60
R20	.57		.26[a]		.61	.39
R21	.66		.33[a]		.46	.54
R22	.50		.16[a]		.73	.27
R23	.67		.11[a]		.54	.46
R24	.66		.02		.56	.44
R25	.60		.02		.64	.36
R26	.57		.11[a]		.66	.34
R27	.57			.06	.67	.33
R28	.58			.04	.66	.34
R29	.53			.25[a]	.66	.34
R30	.42			.13[a]	.81	.19
R31	.73			.23[a]	.41	.59
R32	.32			.26[a]	.83	.17
R33	.62			.24[a]	.56	.44
R34	.39			.35[a]	.73	.27
R35	.40			.31[a]	.74	.26
R36	.54			.39[a]	.55	.45
R37	.44			.42[a]	.63	.37
R38	.40			.31[a]	.74	.26
R39	.64			.37[a]	.46	.54

Note. SMR = squared multiple correlation.

[a] $|t| > 1.96$.

Table 4

Standardized Parameter Estimates for the Single Trait Model (Reading)

Item	Trait	Error [a]	SMR
	Reading [a]		
R1	.52	.73	.27
R2	.56	.69	.31
R3	.66	.57	.43
R4	.66	.56	.44
R5	.46	.78	.22
R6	.58	.66	.34
R7	.47	.78	.22
R9	.71	.50	.50
R10	.31	.90	.10
R11	.51	.74	.26
R12	.72	.48	.52
R13	.55	.70	.30
R14	.66	.57	.43
R15	.38	.85	.15
R16	.64	.59	.41
R17	.60	.64	.36
R18	.53	.72	.28
R19	.75	.44	.56
R20	.57	.67	.33
R21	.66	.56	.44
R22	.50	.75	.25
R23	.68	.54	.46
R24	.66	.56	.44
R25	.59	.56	.35
R26	.58	.67	.33
R27	.58	.67	.33
R28	.59	.65	.35
R29	.55	.69	.31
R30	.44	.81	.19
R31	.76	.42	.58
R32	.35	.88	.12
R33	.64	.59	.41
R34	.43	.81	.19
R35	.44	.81	.19
R36	.58	.66	.34
R37	.49	.76	.24
R38	.44	.81	.19
R39	.67	.55	.45

Note. SMR = squared multiple correlation.

[a] $|t| > 1.96$.

Table 5

Standardized Parameter Estimates for the Single Trait/Correlated Uniqueness Model—Set 1 (Reading)

Item	Trait	Residual covariances											Error	SMR
	Reading	R01	R02	R03	R04	R05	R06	R07	R09	R10	R11	R12		
R1	.53[a]												.72[a]	.28
R2	.55[a]	-.03											.69[a]	.31
R3	.66[a]	.06[a]	-.01										.57[a]	.43
R4	.66[a]	-.04	.01	.02									.56[a]	.44
R5	.46[a]	.01	-.01	.05	-.02								.79[a]	.21
R6	.58[a]	.07[a]	.02	.08[a]	.03	-.08[a]							.67[a]	.33
R7	.47[a]	.02	.04	.06[a]	.05	.01	-.03						.78[a]	.22
R9	.72[a]	-.02	.04	-.04	.03	-.02	-.02	-.04					.49[a]	.51
R10	.30[a]	-.02	.04	.05	.04	.01	.08[a]	.02	.01				.91[a]	.09
R11	.53[a]	.00	.00	-.07[a]	-.04	.00	.04	-.02	-.03	-.01			.72[a]	.28
R12	.71[a]	-.03	.07[a]	-.01	.01	.07[a]	-.02	.00	.09[a]	.00	.03		.49[a]	.51

Note. SMR = squared multiple correlation.

[a] $|t| > 1.96$.

Table 6

Standardized Parameter Estimates for the Single Trait/Correlated Uniqueness Model—Set 2 (Reading)

Item	Trait	Residual covariances														Error	SMR
	Reading	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26		
R13	.57[a]															.68[a]	.32
R14	.64[a]	.10[a]														.58[a]	.42
R15	.40[a]	.09[a]	.09[a]													.84[a]	.16
R16	.67[a]	-.03	-.02	-.02												.56[a]	.44
R17	.62[a]	.00	-.10[a]	-.03	-.05[a]											.62[a]	.38
R18	.55[a]	-.12[a]	-.06	-.04	.08[a]	.04										.70[a]	.30
R19	.77[a]	-.02	-.01	-.09[a]	.03	.06[a]	.01									.40[a]	.60
R20	.57[a]	.00	.13[a]	-.01	-.04	-.09[a]	-.04	-.10[a]								.67[a]	.33
R21	.66[a]	.06	.12[a]	.02	-.02	-.11[a]	-.15[a]	-.02	.04							.57[a]	.43
R22	.48[a]	.00	.07	-.03	-.05	.00	.02	-.02	.03	.12[a]						.77[a]	.23
R23	.67[a]	-.02	.05	-.03	-.03	-.01	-.01	-.03	.12[a]	.02	.04					.55[a]	.45
R24	.67[a]	-.09[a]	.01	-.04	-.03	-.04	.04	-.02	.03	.02	.04	.03				.56[a]	.44
R25	.59[a]	-.02	.07	-.01	.00	.10[a]	.01	-.02	-.01	.03	.03	.00	.04			.66[a]	.34
R26	.58[a]	-.01	.04	.04	-.10[a]	.04	-.06[a]	-.06[a]	-.01	.03	.10[a]	.02	.01	.02			.34

Note. SMR = squared multiple correlation.

[a] |*t*| > 1.96

Table 7

Standardized Parameter Estimates for the Single Trait/Correlated Uniqueness Model—Set 3 (Reading)

Item	Trait	Residual covariances													Error	SMR
	Reading	R27	R28	R29	R30	R31	R32	R33	R34	R35	R36	R37	R38	R39		
R27	.58[a]														.67[a]	.33
R28	.58[a]	.00													.66[a]	.34
R29	.53[a]	-.03	-.06[a]												.72[a]	.28
R30	.42[a]	-.05	.08[a]	.07[a]											.83[a]	.17
R31	.73[a]	.05[a]	.07[a]	.05	.11[a]										.47[a]	.53
R32	.33[a]	.00	.00	.18[a]	.00	-.04									.89[a]	.11
R33	.62[a]	.00	.03	-.02	.03	.09[a]	.02								.62[a]	.38
R34	.40[a]	.02	-.02	.04	.06	.08[a]	.09[a]	.09[a]							.84[a]	.16
R35	.40[a]	.04	.02	.14[a]	.06	.08[a]	.06[a]	.06[a]	.14[a]						.84[a]	.16
R36	.54[a]	.04	.02	.04	.02	.15[a]	.07[a]	.13[a]	.13[a]	.08[a]					.71[a]	.29
R37	.44[a]	-.02	-.01	.22[a]	.05	.07[a]	.08[a]	.09[a]	.17[a]	.15[a]	.15[a]				.81[a]	.19
R38	.40[a]	.02	.04	.02	.01	.06[a]	.10[a]	.05[a]	.08[a]	.11[a]	.13[a]	.14[a]			.84[a]	.16
R39	.64[a]	.04	.02	.07[a]	.03	.06[a]	.16[a]	.10[a]	.10[a]	.09[a]	.16[a]	.12[a]	.15[a]		.56[a]	.41

Note. SMR = squared multiple correlation.

[a] $|t| > 1.96$.

Table 8

Summary of Confirmatory Factor Analysis Model Testing for the Reading Section

Model	Model *df*	S-B scaled chisq	S-B scaled chisq/*df*	GFI	NNFI	CFI	RMSEA 90% CI *p*-value	ECVI 90% CI
Correlated trait & correlated item set[a]	621	1,310.97	2.11	.89	.97	.97	.020 .019–.022 1.00	.57 [b] .53–-.61
Correlated trait & uncorrelated item set	624	1,383.56	2.22	.89	.97	.97	.021 .020–.023 1.00	.59[b] .56–.64
Correlated trait	662	2,068.42	3.12	.84	.96	.96	.028 .027–.029 1.00	.82[b] .77–.87
Correlated trait & correlated uniqueness	438	846.74	1.93	.92	.97	.98	.019 .017–.020 1.00	.53 .51–.57
Single trait & correlated item set[a]	624	1,349.49	2.16	.89	.97	.97	.021 .019–.022 1.00	.58 [b] .54–.62
Single trait & uncorrelated item set	627	1,400.02	2.23	.89	.97	.97	.021 .020–.023 1.00	.60[b] .56–.64

(Table continues)

Table 8 (continued)

Model	Model *df*	S-B scaled chisq	S-B scaled chisq/*df*	GFI	NNFI	CFI	RMSEA 90% CI *p*-value	ECVI 90% CI
Single trait	665	2,076.99	3.12	.84	.96	.96	.028 .027–.029 1.00	.82[b] .77–.87
Single trait & correlated uniqueness	441	855.57	1.94	.92	.97	.98	.019 .017–.020 1.00	.54 .51–.57

Note. S-B scaled chisq = Satorra-Bentler scaled chi-square; GFI = goodness of fit index; NNFI = non-normed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; ECVI = expected cross-validation index; CI = confidence interval.

[a] The model parameter estimates were not substantively interpretable. [b] The estimated ECVI was larger than that of the saturated model.

The single trait/uncorrelated item set and single trait/correlated uniqueness model could not be compared by means of a chi-square difference test, because these models are not nested with each other. However, the relatively better fit of the correlated trait/correlated uniqueness model suggests that the method effects present in the Reading section were not unidimensional.

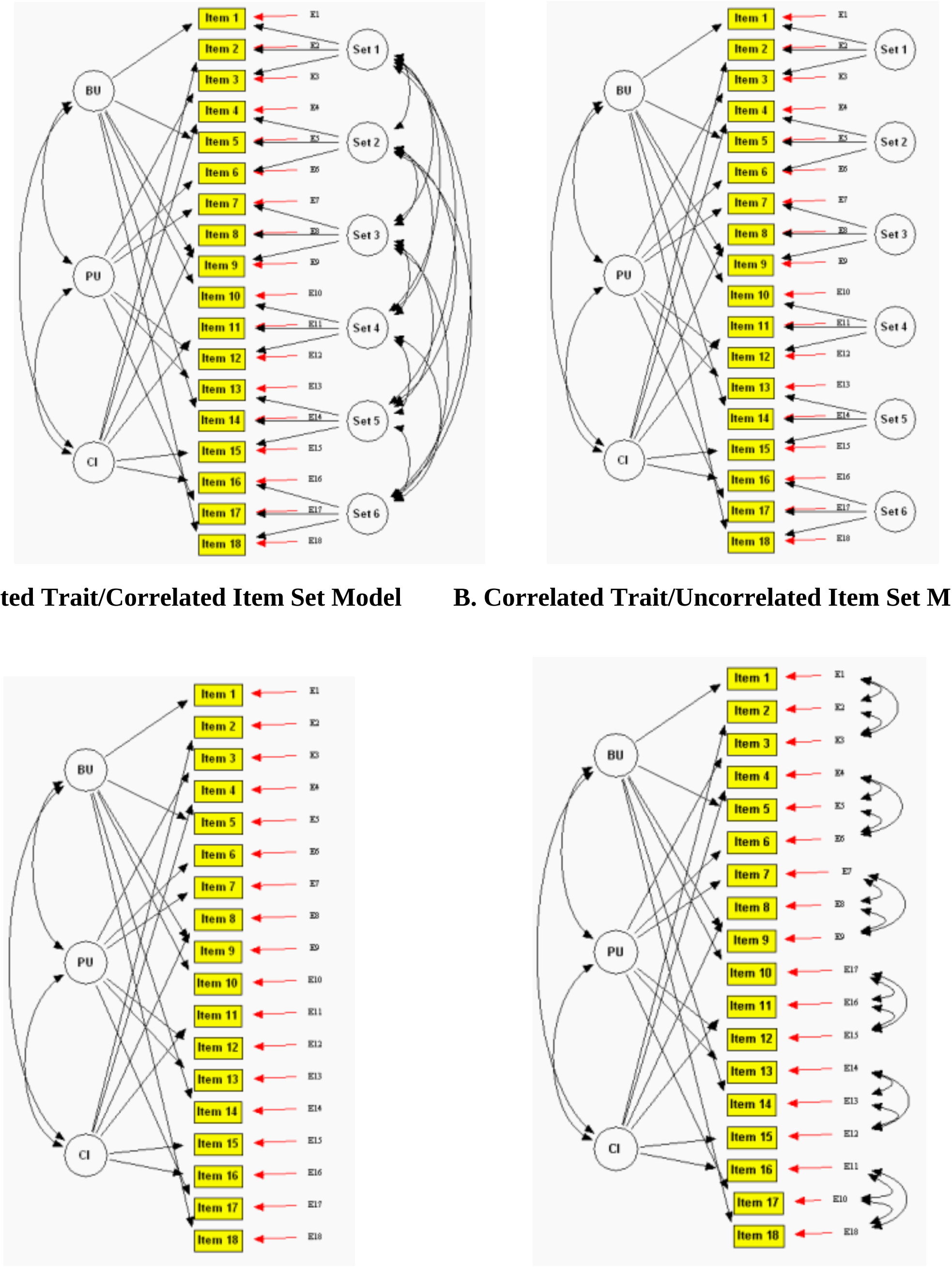
Finally, the parameter estimates were compared across the three models. The standardized model parameter estimates for these models are presented in Tables 3-7. The magnitudes of the trait factor loadings were remarkably similar, a majority being identical across the three models. When the trait factor loadings were different across the models, the fluctuations were within ± .05. The item set effects in the single trait/uncorrelated item set and single trait/correlated uniqueness models were most pronounced for Set 3. In the single trait/uncorrelated item set model, 11 out of the 13 items were associated with significant low to moderate item set factor loadings, and in the single trait/correlated uniqueness model, the majority of the residual correlations among the items in Set 3 were significant. The pattern of the significant residual correlations indicated the item set effects on individual items, providing some evidence that the item set effects were not uniform across the items associated with an item set.

Taken together, the generally good fit of the single trait/correlated uniqueness model and the single trait/uncorrelated item set model suggested the presence of item set effects as defined by the item sets. Between these two models, the relatively better goodness-of-fit indices for the single trait/correlated uniqueness model further suggested the possibility that the item set effects were not uniform across the individual items. Despite the indication of the item set effects, however, the fit of the single trait model was still marginally acceptable when considering the practical criteria of goodness of fit of the model. In sum, these results suggest that, although potentially non-unidimensional item set effects were present in the Reading data structure, the effects were not so pronounced as to be practically important. Furthermore, the stability of the model parameter estimates across the single trait model and the other two models that took account of the item set effects indicated that the interpretation of the trait factor structure was not affected by whether or not the item set effects were explicitly modeled. For these reasons, the single trait model was adopted as a parsimonious representation of the trait factor structure of the Reading section. The adoption of the model with a single trait factor also suggested the unidimensional nature of the reading abilities assessed in the Reading section.

Listening section. The analysis of the Listening section was parallel to that of the Reading section. All MTMM models for the Listening section specified three trait factors corresponding to the three-part definitions of the target constructs as defined in the test specifications: Basic Understanding, Pragmatic Understanding, and Connecting Information. The six item-set factors corresponded to the six item sets, the first two based on conversations, and the latter four on academic lectures. Initially, the same four alternative MTMM models with correlated traits as those tested for the Reading section—correlated trait/correlated item set, correlated trait/uncorrelated item set, correlated trait, and correlated trait/correlated uniqueness—were tested (see Figure 2 for sample schematics).

All four models converged. Although all of them produced moderate trait factor loadings and fairly small item set factor loadings, the model parameter estimates for the correlated trait/correlated item set model tended to deviate considerably from those of the others. There were two problems with these models. First, the trait-factor correlations in all the models were extremely high (over .98). Second, the correlated trait/uncorrelated item set and the correlated trait models indicated some model estimation problems, both models producing out-of-range parameter estimates.

Because all these models suggested the lack of psychometric distinctiveness among the three traits specified, the single trait counterparts of these models (single trait/correlated item set, single trait/uncorrelated item set, single trait, and single trait/correlated uniqueness models) were developed by fixing all the trait-factor correlations in the initial models to 1.0. Among the four, the single trait/correlated item set model converged but with uninterpretable factor loading patters, featuring negative low to moderate trait factor loadings and positive moderate method factor loadings. Moreover, the single trait/uncorrelated item set model did not converge. Thus, these models were not considered further. These results indicate that the item set effects in the Listening section may not be unidimensional across the items. The single trait and single trait/correlated uniqueness models converged with reasonable parameter estimates and standard errors (see Tables 9–16 for the model parameter estimates and the model fit indices). Because the single trait and single trait/correlated uniqueness models showed reasonable fit to the data, further comparisons were made between these two models.



A. Correlated Trait/Correlated Item Set Model

B. Correlated Trait/Uncorrelated Item Set Model

C. Correlated Trait Model

D. Correlated Trait/Correlated Uniqueness Model

***Figure 2*. Schematic representation of the four initial MTMM models tested forthe Listening section.**

Note. BU = Basic Understanding; PU = Pragmatic Understanding; CI = Connecting Information.

Table 9

Standardized Parameter Estimates for Single Trait Model (Listening)

Item	Trait	Error	SMR
	Listening		
L1	.60[a]	.65[a]	.35
L2	.67[a]	.55[a]	.45
L3	.70[a]	.50[a]	.50
L4	.53[a]	.72[a]	.28
L5	.65[a]	.58[a]	.42
L6	.45[a]	.79[a]	.21
L7	.52[a]	.73[a]	.27
L8	.58[a]	.66[a]	.34
L9	.63[a]	.60[a]	.40
L10	.77[a]	.41[a]	.59
L11	.37[a]	.86[a]	.14
L12	.59[a]	.65[a]	.35
L13	.61[a]	.63[a]	*.37*
L14	.76[a]	.42[a]	.58
L15	.54[a]	.71[a]	.29
L16	.51[a]	.74[a]	.26
L17	.64[a]	.59[a]	.41
L18	.51[a]	.74[a]	.26
L19	.53[a]	.72[a]	.28
L20	.70[a]	.51[a]	.49
L21	.74[a]	.45[a]	.55
L22	.53[a]	.72[a]	.28
L23	.66[a]	.57[a]	.43
L24	.72[a]	.48[a]	.52
L25	.78[a]	.40[a]	.60
L26	.70[a]	.51[a]	.49
L27	.74[a]	.46[a]	.54
L28	.65[a]	.58[a]	.42
L29	.51[a]	.74[a]	.26
L30	.50[a]	.75[a]	.25
L31	.60[a]	.63[a]	.37
L32	.39[a]	.85[a]	.15
L34	.35[a]	.88[a]	.12

Note. SMR = Squared multiple correlation.

[a] $|t| > 1.96$.

Table 10

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 1 (Listening)

Item	Trait	Residual covariances					Error	SMR
	Listening	L01	L02	L03	L04	L05		
L1	.59[a]	.66[a]					.66[a]	.34
L2	.66[a]	.08[a]	.56[a]				.56[a]	.44
L3	.70[a]	.11[a]	.06[a]	.51[a]			.51[a]	.49
L4	.51[a]	.10[a]	.13[a]	.10[a]	.74[a]		.74[a]	.26
L5	.64[a]	.07[a]	.11[a]	.01	.03	.59[a]	.59[a]	.41

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 11

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 2 (Listening)

Item	Trait	Residual covariances					Error	SMR[a]
	Listening	L06	L07	L08	L09	L10		
L6	.45[a]	.80[a]					.80[a]	.20
L7	.51[a]	-.02	.74[a]				.74[a]	.26
L8	.58[a]	.04	.02	.66[a]			.66[a]	.34
L9	.63[a]	.03	.03	.04	.60[a]		.60[a]	.40
L10	.77[a]	.07[a]	.09[a]	.03	.04	.41[a]	.41[a]	.59

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 12

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 3 (Listening)

Item	Trait	Item						Trait	SMR
	Listening	L11	L12	L13	L14	L15	L16		
L11	.36[a]	.87[a]						.87[a]	.13
L12	.58[a]	.04	.67[a]					.67[a]	.33
L13	.60[a]	.08[a]	.08[a]	.64[a]				.64[a]	.36
L14	.75[a]	.09[a]	.17[a]	.12[a]	.43[a]			.43[a]	.57
L15	.53[a]	.03	.05	.07[a]	.06[a]	.72[a]		.72[a]	.28
L16	.52[a]	.01	-.04	.02	-.04	-.08[a]	.73[a]	.73[a]	.27

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 13

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 4 (Listening)

Item	Trait	Residual covariances						Error	SMR
	Listening	L17	L18	L19	L20	L21	L22		
L17	.64[a]	.59[a]						.59[a]	.41
L18	.51[a]	.05	.74[a]					.74[a]	.26
L19	.53[a]	.02	.11[a]	.72[a]				.72[a]	.28
L20	.70[a]	-.01	.04	-.02	.51[a]			.51[a]	.49
L21	.74[a]	.03	.00	.00	.07[a]	.45[a]		.45[a]	.55
L22	.54[a]	.02	-.03	.02	-.07[a]	-.02	.71[a]	.71[a]	.29

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 14

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 5 (Listening)

Item	Trait	Residual covariances						Error	SMR
	Listening	L23	L22	L25	L26	L27	L28		
L23	.64[a]	.59[a]						.59[a]	.41
L24	.71[a]	.10[a]	.50[a]					.50[a]	.50
L25	.76[a]	.04	.07[a]	.42[a]				.42[a]	.58
L26	.68[a]	.07[a]	.07[a]	.05	.54[a]			.54[a]	.46
L27	.73[a]	.05	.04	.07[a]	.06[a]	.47[a]		.47[a]	.53
L28	.63[a]	.09[a]	.04	.09[a]	.17[a]	.08[a]	.61[a]	.61[a]	.39

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 15

Standardized Parameter Estimates for Single Trait/Correlated Uniqueness Model—Set 6 (Listening)

Item	Trait	Residual covariances					Error	SMR
	Listening	L29	L30	L31	L32	L34		
L29	.51[a]	.74[a]					.74[a]	.26
L30	.50[a]	.05	.75[a]				.75[a]	.25
L31	.60[a]	.05[a]	.04	.64[a]			.64[a]	.36
L32	.39[a]	.04	.12[a]	.10[a]	.84[a]		.85[a]	.15
L34	.35[a]	.02	-.05	.04	-.05	.88[a]	.88[a]	.12

Note. SMR = Squared multiple correlation.

[a] |*t*| > 1.96.

Table 16

Summary of Confirmatory Factor Analysis Model Testing for the Listening Section

Model	Model *df*	S-B scaled chisq	S-B scaled chisq/*df*	GFI	NNFI	CFI	RMSEA 90% CI[d]	ECVI 90% CI
Correlated trait & correlated item set	444	579.64	1.30	.94	.98	.99	.011 .008–.013	.30 .28–.32
Correlated trait & uncorrelated item set[a]	459	784.52	1.71	.92	.98	.98	.016 .014–.018	.36 .34–.39
Correlated trait[b]	492	1,346.06	2.74	.87	.97	.97	.025 .024–.027	.55 [e] .51–.59
Correlated trait & correlated uniqueness	417	681.59	1.63	.93	.98	.98	.015 .013–.017	.36 .33–.38
Single trait & correlated item set[c]	447	562.11	1.28	.94	.98	.99	.010 .007–.012	.29 .27–.31
Single trait & uncorrelated item set				No convergence				
Single trait	495	1,348.42	2.72	.87	.97	.97	.025 .024–.027	.54 [e] .51–.59
Single trait & correlated uniqueness	420	681.48	1.62	.93	.98	.98	.015 .013–.017	.35 .33–.38

Note. S-B = Satorra-Bentler scaled chi-square; GFI = goodness of fit index; NNFI = non-normed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; ECVI = expected cross-validation index; CI = confidence interval.

[a] Four parameter estimates were out of range. [b] One interfactor correlation was out of range. [c] The model parameter estimates were not interpretable. [d] *p*-value = 1.00. [e] The estimated ECVI was larger than that of the saturated model.

The goodness-of-fit indices for the single trait and single trait/correlated uniqueness models shown in Table 16 indicate adequate fit of both models on all the criteria except for the slightly lower GFI (.87) for the single trait model, as well as the significant Satorra-Bentler model chi-square values for both models. A chi-square difference test showed that the fit of the single trait/correlated uniqueness model was significantly better than that of the single trait model ($p < .01$; $\chi^2_{\text{S-B difference}} = 932.92$; $df = 75$). In addition, the ECVI value for the single trait/correlated uniqueness model was smaller, indicating a better chance of replicating the same result in a different sample. However, considering the indices of model fit, the practical improvement of the model fit by modeling the correlated uniqueness in the single trait/correlated uniqueness model was marginal, with the exceptions of GFI and ECVI. Moreover, the trait factor loadings for the single trait and single trait/correlated uniqueness models (see Tables 9–16) were roughly identical. For this reason, the single trait model was adopted as a parsimonious representation of the internal structure of the Listening section.

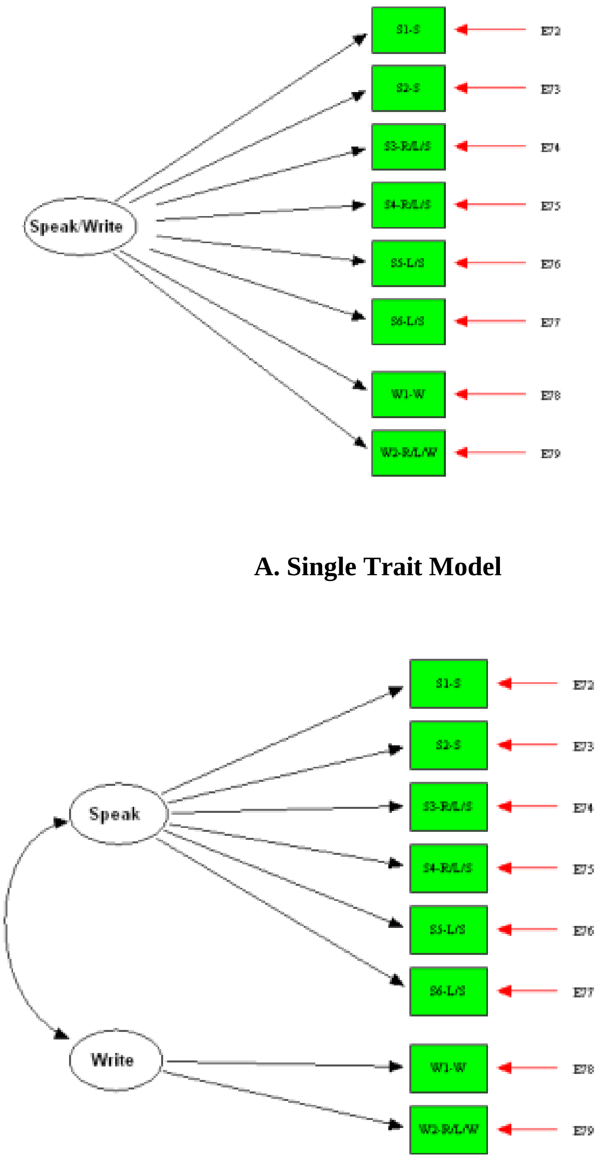
To sum up the results of the MTMM analysis of the Listening section, the equivalent goodness of the fit of the single trait models to their correlated trait counterparts suggested the psychometric unidimensionality of the three traits as defined in the test specifications. Moreover, the estimation problems associated with the single trait/correlated method and single trait/uncorrelated method models as opposed to the proper and interpretable results of the single trait/correlated uniqueness model indicate that the item set effects present in the Listening section were not unidimensional, suggesting that the individual items were affected by the method effects to different degrees. Finally, the practically equivalent fit of the single trait and single trait/correlated uniqueness models showed that, although some method effects were present in the Listening section, they were not pronounced.

Speaking and Writing sections. Another series of analyses was conducted for the item responses for the Speaking and Writing sections. Because there were only eight measured variables in the Speaking and Writing sections in total, the data from the two sections were analyzed together. Two CFA models with only one or two factors presented in Figure 3 were tested at this stage of the analyses. (A separate CFA model for the Writing section could not be tested because there were only two items in the section—testing a CFA model with only two measured variables will result in model identification problems). One model tested was a single factor model that specified only one factor (i.e., the Speaking/Writing factor) on which all the speaking and writing items loaded. The other model was a correlated two-factor model. This model specified one factor for each modality (i.e., the Speaking and Writing factors), so that the loadings of the six speaking items on the Speaking factor and those of the two writing items on the Writing factor were freely estimated. The covariance between the Speaking and Writing factors was estimated freely as well.

The parameter estimates and the goodness-of-fit measures for these two models are presented in Tables 17 and 18. The factor loadings of the items were highly similar across the two models for the Speaking section, while those for the Writing items were less stable, with more variation across the two models. The interfactor correlation obtained for the correlated trait model was .86, which was considerably high but still suggested the distinctness of the traits according to the predetermined rule of thumb of .90. The Satorra-Bentler model chi-square and $\chi^2_{\text{S-B}}$/df were large for both models, partially due to the large sample size as opposed to the small degrees of freedom associated with these models. Although the goodness of fit of both the single trait and correlated trait models were acceptable in terms of the GFI, NNFI, and CFI, the fit of the single trait model was poorer than that of the correlated trait model in terms of the $\chi^2_{\text{S-B}}$/df and RMSEA values. Moreover, the smaller ECVI for the correlated trait model indicated that the correlated trait model would replicate better in a different sample. A chi-square difference test also suggested a significantly better fit of the correlated trait model ($p < .01$; $\chi^2_{\text{S-B difference}} = 242.34$; $df = 1$). Collectively, these findings supported the correlated trait model as the better representation of the structure of the combined Speaking and Writing sections.

Confirmatory Factor Analyses of the Entire Test

The single trait models adopted for the Reading and Listening sections separately and the correlated two-factor model for the Speaking and Writing sections were combined in order to develop a CFA model for the entire test. As already noted, the TOEFL iBT Speaking and Writing sections involve integrated tasks. Thus, two series of models were tested in the subsequent analyses of the entire test by alternating the way in which the integrated tasks were specified in the CFA models. In one series of the analyses, all CFA models estimated the loading of each integrated task on the factor representing the target modality only, consistent with the test design (e.g., only the loading of the Listening/Speaking task on the Speaking factor was estimated, while that on the Listening factor was not; see Table 19 for the fit indices for these models). The second series of the analyses allowed cross-loadings of the integrated tasks to more than one factor, so that the loadings of each integrated task on all the associated modalities were estimated (e.g., the loadings of the Listening/Speaking task on both the Speaking and Listening factors were estimated).



A. Single Trait Model

B. Correlated Two-Factor Model

***Figure 3*. Two alternative confirmatory factor analysis models tested for the Speaking and Writing sections.**

Table 17

Standardized Parameter Estimates (Speaking and Writing)

Item	Single factor model			Correlated trait model			
	S/W [a]	Error [a]	SMR	S [a]	W [a]	Error [a]	SMR
S1 (S)	.77	.41	.59	.77		.40	.60
S2 (S)	.77	.40	.60	.78		.40	.60
S3 (R/L/S)	.84	.30	.70	.84		.30	.70
S4 (R/L/S)	.86	.25	.75	.87		.25	.75
S5 (L/S)	.86	.27	.73	.86		.26	.74
S6 (L/S)	.82	.32	.68	.82		.32	.68
W1 (W)	.78	.39	.61		.89	.21	.79
W2 (R/L/W)	.72	.49	.51		.81	.35	.65
				Trait-factor correlations			
S				1.00			
W				.86[a]	1.00		

Note. S = Speaking; R = Reading; L = Listening; S/W = Speaking/Writing; S = Speaking; W = Writing; SMR = squared multiple correlation.

[a] |*t*| > 1.96.

Table 18

Summary of Confirmatory Factor Analysis Model Testing for the Speaking and Writing Sections

Model	Model *df*	S-B scaled chisq	S-B scaled chisq/ *df*	GFI	NNFI	CFI	RMSEA 90% CI *p*-value	ECVI (90% CI)
Single trait model	20	369.95	18.45	.94	.97	.98	.080 .073–.087 0.00	.15[a] (.13–.17)
Correlated two-factor model	19	119.71	6.30	.98	.99	.99	.044 .037–.052 .89	.06[a] (.05–.07)

Note. S-B scaled chisq= Satorra-Bentler scaled chi-square; GFI = goodness of fit index; NNFI = non-normed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; ECVI = expected cross-validation index; CI = confidence interval.

[a] The estimated ECVI was larger than that of the saturated model.

Table 19

Summary of Confirmatory Factor Analysis Model Testing for the Entire Test Without Additional Paths for the Integrated Tasks (All Sections)

Model	Model *df*	S-B scaled chisq	S-B scaled chisq/ *df*	GFI	NNFI	CFI	RMSEA 90% CI[a]	ECVI 90% CI
Bifactor	2,917	5,188.33	1.78	.82	.98	.98	.017 .016–.018	2.09 2.01–2.16
Correlated traits	2,996	6,780.85	2.26	.78	.98	.98	.022 .021–.022	2.61[b] 2.53–2.70
Single trait	3,002	11,314.98	3.78	.69	.97	.97	.032 .031–.033	4.28[b] 4.16–4.40
Higher-order factor	2,998	6,893.62	2.30	.78	.98	.98	.022 .021–.023	2.66[b] 2.57–2.74

Note. S-B scaled chisq = Satorra-Bentler scaled chi-square; GFI = goodness of fit index; NNFI = non-normed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; ECVI = expected cross-validation index; CI = confidence interval.
[a] *p*-value = 1.00. [b] The estimated ECVI was larger than that of the saturated model.

One problem in the second series of the analyses was that the loadings of the Reading/Listening/Writing task on all the three associated modalities—Reading, Listening, and Writing—could not be modeled simultaneously in a CFA model. Because the Writing section contained only two items, the construct assessed in this section was not sufficiently measured to create a latent variable. Primarily for this reason, it was determined a priori that freely estimating all the paths associated with the independent and integrated Writing tasks would result in an unidentified model. After attempting various combinations of the paths to be specified for the integrated tasks, CFA models that estimated full paths for all four integrated Speaking tasks (two paths for the Listening/Speaking tasks and three paths for the Reading/Listening/Speaking tasks) but only two paths for the integrated Writing task (the paths for Reading and Writing only, while dropping the path for Listening) converged with proper parameter estimates (see Tables 20–23 for the parameter estimates and Table 24 for the model fit indices).

Table 20

Standardized Parameter Estimates for the Bifactor Model With Additional Paths for the Integrated Tasks (All Sections)

Item	General	Reading[a]	Listening[a]	Speaking[a]	Writing[a]	Error[a]	SMR
R1	-.14	.52				.70	.30
R2	.00	.56				.69	.31
R3	.04	.66				.56	.44
R4	.07	.67				.55	.45
R5	-.04	.46				.78	.22
R6	.01	.57				.67	.33
R7	.01	.47				.78	.22
R9	.08	.71				.49	.51
R10	.15[a]	.30				.89	.11
R11	-.19[a]	.51				.70	.30
R12	.04	.72				.48	.52
R13	.19[a]	.54				.67	.33
R14	.40[a]	.66				.41	.59
R15	.11	.37				.85	.15
R16	-.12	.66				.55	.45
R17	-.15	.61				.61	.39
R18	-.20[a]	.54				.67	.33
R19	-.14	.76				.40	.60
R20	.27[a]	.57				.60	.40
R21	.29[a]	.66				.49	.51
R22	.14	.49				.74	.26
R23	.15	.68				.51	.49
R24	.04	.66				.57	.43
R25	.03	.59				.65	.35
R26	.06	.57				.67	.33
R27	-.13	.59				.64	.36
R28	.04	.58				.66	.34
R29	.06	.55				.69	.31
R30	.04	.43				.82	.18
R31	.07	.75				.43	.57
R32	-.25[a]	.35				.82	.18
R33	-.10	.64				.58	.42
R34	-.05	.43				.81	.19
R35	-.05	.44				.80	.20

(Table continues)

Table 20 (continued)

Item	General	Reading[a]	Listening[a]	Speaking[a]	Writing[a]	Error[a]	SMR
R36	-.10	.58				.65	.35
R37	.04	.48				.77	.23
R38	-.08	.43				.81	.19
R39	-.21[a]	.68				.50	.50
L1	.08		.60			.63	.37
L2	-.20		.67			.51	.49
L3	.03		.70			.51	.49
L4	.02		.52			.73	.27
L5	-.15		.64			.57	.43
L6	-.19[a]		.45			.76	.24
L7	.16		.52			.70	.30
L8	.00		.58			.66	.34
L9	-.05		.64			.59	.41
L10	.04		.77			.40	.60
L11	-.04		.37			.86	.14
L12	-.27[a]		.58			.59	.41
L13	-.20[a]		.61			.59	.41
L14	-.23[a]		.75			.38	.62
L15	-.21[a]		.54			.67	.33
L16	.10		.52			.72	.28
L17	.08		.64			.59	.41
L18	-.12		.51			.72	.28
L19	-.05		.53			.72	.28
L20	-.24[a]		.70			.46	.54
L21	-.04		.74			.45	.55
L22	.23[a]		.54			.66	.34
L23	.11		.66			.55	.45
L24	.07		.73			.47	.53
L25	.14		.78			.37	.63
L26	.18		.71			.46	.54
L27	.13		.74			.44	.56
L28	.13		.65			.56	.44
L29	-.01		.51			.73	.27
L30	-.19[a]		.50			.71	.29
L31	.01		.60			.64	.36
L32	-.22[a]		.39			.80	.20
L34	.20[a]		.35			.84	.16

(Table continues)

Table 20 (continued)

Item	General	Reading[a]	Listening[a]	Speaking[a]	Writing[a]	Error[a]	SMR
S1 (S)	-.04			.78		.39	.61
S2 (S)	-.08			.78		.39	.61
S3 (R/L/S)	-.08	.06	.14	.68		.30	.70
S4 (R/L/S)	-.09	-.02	.15	.76		.25	.75
S5 (L/S)	-.02		.08	.81		.25	.75
S6 (L/S)	-.04		.13	.72		.32	.68
W1 (W)	-.14				.89	.18	.82
W2 (R/L/W)	-.04	-.04			.83	.37	.63
Interfactor correlations							
R		1.00					
L		.89[a]	1.00				
S		.66[a]	.76[a]	1.00			
W		.87[a]	.89[a]	.82[a]	1.00		

Note. G = general; R = Reading; L = Listening; S = Speaking; W = Writing; SMR = squared multiple correlations.

[a] |t|> 1.96.

Table 21

Standardized Parameter Estimates for the Correlated Trait Model With Additional Paths for the Integrated Tasks (All Sections)

Item	Reading	Listening	Speaking	Writing	Error	SMR
R1	.52[a]				.73[a]	.27
R2	.56[a]				.68[a]	.32
R3	.66[a]				.56[a]	.44
R4	.67[a]				.56[a]	.44
R5	.46[a]				.79[a]	.21
R6	.57[a]				.67[a]	.33
R7	.47[a]				.78[a]	.22
R9	.71[a]				.49[a]	.51
R10	.30[a]				.91[a]	.09
R11	.51[a]				.74[a]	.26
R12	.72[a]				.48[a]	.52
R13	.54[a]				.70[a]	.30

(Table continues)

Table 21 (continues)

Item	Reading	Listening	Speaking	Writing	Error	SMR
R14	.65[a]				.57[a]	.43
R15	.38[a]				.86[a]	.14
R16	.65[a]				.57[a]	.43
R17	.60[a]				.63[a]	.37
R18	.53[a]				.72[a]	.28
R19	.76[a]				.42[a]	.58
R20	.57[a]				.67[a]	.33
R21	.66[a]				.57[a]	.43
R22	.49[a]				.76[a]	.24
R23	.69[a]				.53[a]	.47
R24	.66[a]				.57[a]	.43
R25	.59[a]				.65[a]	.35
R26	.57[a]				.67[a]	.33
R27	.58[a]				.66[a]	.34
R28	.58[a]				.66[a]	.34
R29	.55[a]				.70[a]	.30
R30	.43[a]				.82[a]	.18
R31	.75[a]				.43[a]	.57
R32	.34[a]				.88[a]	.12
R33	.64[a]				.59[a]	.41
R34	.43[a]				.82[a]	.18
R35	.44[a]				.81[a]	.19
R36	.58[a]				.66[a]	.34
R37	.48[a]				.77[a]	.23
R38	.43[a]				.81[a]	.19
R39	.67[a]				.55[a]	.45
L1		.60[a]			.64[a]	.36
L2		.68[a]			.54[a]	.46
L3		.70[a]			.51[a]	.49
L4		.52[a]			.73[a]	.27
L5		.64[a]			.59[a]	.41
L6		.46[a]			.79[a]	.21
L7		.52[a]			.73[a]	.27
L8		.58[a]			.66[a]	.34
L9		.64[a]			.59[a]	.41
L10		.77[a]			.41[a]	.59
L11		.37[a]			.86[a]	.14
L12		.59[a]			.66[a]	.34
L13		.61[a]			.63[a]	.37
L14		.76[a]			.43[a]	.57
L15		.54[a]			.71[a]	.29

(Table continues)

Table 21 (continues)

Item	Reading	Listening	Speaking	Writing	Error	SMR
L16		.52[a]			.73[a]	.27
L17		.63[a]			.60[a]	.40
L18		.52[a]			.73[a]	.27
L19		.53[a]			.72[a]	.28
L20		.70[a]			.51[a]	.49
L21		.74[a]			.45[a]	.55
L22		.53[a]			.72[a]	.28
L23		.66[a]			.57[a]	.43
L24		.73[a]			.47[a]	.53
L25		.78[a]			.40[a]	.60
L26		.71[a]			.50[a]	.50
L27		.73[a]			.46[a]	.54
L28		.65[a]			.58[a]	.42
L29		.52[a]			.73[a]	.27
L30		.50[a]			.75[a]	.25
L31		.60[a]			.64[a]	.36
L32		.40[a]			.84[a]	.16
L34		.34[a]			.88[a]	.12
S1 (S)			.78[a]		.39[a]	.61
S2 (S)			.78[a]		.39[a]	.61
S3 (R/L/S)	.05	.15[a]	.68[a]		.30[a]	.70
S4 (R/L/S)	-.03	.15[a]	.77[a]		.25[a]	.75
S5 (L/S)		.08[a]	.80[a]		.26[a]	.74
S6 (L/S)		.13[a]	.72[a]		.33[a]	.67
W1 (W)				.90[a]	.19[a]	.81
W2 (R/L/W)	-.01			.80[a]	.37[a]	.63
Interfactor correlations						
R	1.00					
L	.89[a]	1.00				
S	.66[a]	.76[a]	1.00			
W	.86[a]	.89[a]	.82[a]	1.00		

Note. G = general; R = Reading; L = Listening; S = Speaking; W = Writing; SMR = squared multiple correlations.

[a] |t|> 1.96.

Table 22

Standardized Parameter Estimates for the Single Trait Model With Additional Paths for the Integrated Tasks (All Sections)

Item	General	Error	SMR
R1	.50[a]	.75[a]	.25
R2	.55[a]	.70[a]	.30
R3	.64[a]	.59[a]	.41
R4	.66[a]	.56[a]	.44
R5	.43[a]	.81[a]	.19
R6	.53[a]	.72[a]	.28
R7	.45[a]	.79[a]	.21
R9	.70[a]	.51[a]	.49
R10	.27[a]	.93[a]	.07
R11	.50[a]	.75[a]	.25
R12	.69[a]	.53[a]	.47
R13	.51[a]	.74[a]	.26
R14	.61[a]	.63[a]	.37
R15	.34[a]	.89[a]	.11
R16	.65[a]	.58[a]	.42
R17	.58[a]	.66[a]	.34
R18	.51[a]	.74[a]	.26
R19	.76[a]	.43[a]	.57
R20	.57[a]	.68[a]	.32
R21	.62[a]	.62[a]	.38
R22	.45[a]	.80[a]	.20
R23	.68[a]	.54[a]	.46
R24	.62[a]	.62[a]	.38
R25	.55[a]	.69[a]	.31
R26	.53[a]	.72[a]	.28
R27	.57[a]	.67[a]	.33
R28	.54[a]	.70[a]	.30
R29	.53[a]	.72[a]	.28
R30	.39[a]	.85[a]	.15
R31	.71[a]	.50[a]	.50
R32	.32[a]	.90[a]	.10
R33	.61[a]	.63[a]	.37
R34	.41[a]	.83[a]	.17
R35	.42[a]	.82[a]	.18
R36	.54[a]	.71[a]	.29
R37	.44[a]	.81[a]	.19
R38	.40[a]	.84[a]	.16
R39	.64[a]	.59[a]	.41
L1	.59[a]	.66[a]	.34

(Table continues)

Table 22 (continues)

Item	General	Error	SMR
L2	.66[a]	.56[a]	.44
L3	.67[a]	.55[a]	.45
L4	.50[a]	.75[a]	.25
L5	.62[a]	.61[a]	.39
L6	.45[a]	.80[a]	.20
L7	.50[a]	.75[a]	.25
L8	.57[a]	.68[a]	.32
L9	.63[a]	.61[a]	.39
L10	.75[a]	.44[a]	.56
L11	.37[a]	.86[a]	.14
L12	.56[a]	.68[a]	.32
L13	.60[a]	.64[a]	.36
L14	.73[a]	.46[a]	.54
L15	.53[a]	.72[a]	.28
L16	.52[a]	.73[a]	.27
L17	.62[a]	.62[a]	.38
L18	.51[a]	.74[a]	.26
L19	.52[a]	.72[a]	.28
L20	.69[a]	.53[a]	.47
L21	.73[a]	.47[a]	.53
L22	.53[a]	.72[a]	.28
L23	.64[a]	.59[a]	.41
L24	.71[a]	.49[a]	.51
L25	.76[a]	.43[a]	.57
L26	.70[a]	.50[a]	.50
L27	.70[a]	.50[a]	.50
L28	.63[a]	.60[a]	.40
L29	.51[a]	.74[a]	.26
L30	.50[a]	.75[a]	.25
L31	.59[a]	.65[a]	.35
L32	.40[a]	.84[a]	.16
L34	.34[a]	.89[a]	.11
S1 (S)	.62	.62[a]	.38
S2 (S)	.62	.61[a]	.39
S3 (R/L/S)	.73	.47[a]	.53
S4 (R/L/S)	.72	.48[a]	.52
S5 (L/S)	.71	.50[a]	.50
S6 (L/S)	.70	.51[a]	.49
W1 (W)	.83	.30[a]	.70
W2 (R/L/W)	.73	.47[a]	.53

Note. G = general; SMR = squared multiple correlations.

[a] |t|> 1.96.

Table 23

Standardized Parameter Estimates for the Higher-Order Factor Model With Additional Paths for the Integrated Tasks (All Sections)

Item	Reading	Listening	Speaking	Writing	General	Error	SMR
R1	.52[b]					.73[a]	.27
R2	.56[a]					.68[a]	.32
R3	.66[a]					.56[a]	.44
R4	.67[a]					.56[a]	.45
R5	.46[a]					.79[a]	.21
R6	.57[a]					.67[a]	.33
R7	.47[a]					.78[a]	.22
R9	.71[a]					.49[a]	.51
R10	.30[a]					.91[a]	.09
R11	.51[a]					.74[a]	.26
R12	.72[a]					.48[a]	.52
R13	.55[a]					.70[a]	.30
R14	.65[a]					.57[a]	.43
R15	.38[a]					.86[a]	.14
R16	.65[a]					.57[a]	.43
R17	.60[a]					.64[a]	.36
R18	.53[a]					.72[a]	.28
R19	.76[a]					.42[a]	.58
R20	.58[a]					.67[a]	.33
R21	.66[a]					.57[a]	.43
R22	.49[a]					.76[a]	.24
R23	.69[a]					.53[a]	.47
R24	.66[a]					.57[a]	.43
R25	.59[a]					.65[a]	.35
R26	.57[a]					.67[a]	.33
R27	.58[a]					.66[a]	.34
R28	.58[a]					.66[a]	.34
R29	.55[a]					.70[a]	.30
R30	.43[a]					.82[a]	.18
R31	.75[a]					.43[a]	.57
R32	.34[a]					.88[a]	.12
R33	.64[a]					.59[a]	.41
R34	.43[a]					.81[a]	.19
R35	.44[a]					.81[a]	.19
R36	.58[a]					.66[a]	.34
R37	.48[a]					.77[a]	.23
R38	.43[a]					.81[a]	.19
R39	.67[a]					.55[a]	.45

(Table continues)

Table 23 (continued)

Item	Reading	Listening	Speaking	Writing	General	Error	SMR
L1		.60[b]				.64[a]	.36
L2		.68[a]				.54[a]	.46
L3		.70[a]				.51[a]	.49
L4		.52[a]				.73[a]	.27
L5		.65[a]				.58[a]	.42
L6		.46[a]				.79[a]	.21
L7		.52[a]				.73[a]	.27
L8		.58[a]				.66[a]	.34
L9		.64[a]				.59[a]	.41
L10		.77[a]				.41[a]	.59
L11		.37[a]				.86[a]	.14
L12		.59[a]				.66[a]	.34
L13		.61[a]				.63[a]	.38
L14		.76[a]				.43[a]	.57
L15		.54[a]				.71[a]	.30
L16		.52[a]				.73[a]	.27
L17		.63[a]				.60[a]	.40
L18		.52[a]				.73[a]	.27
L19		.53[a]				.72[a]	.28
L20		.70[a]				.51[a]	.49
L21		.74[a]				.45[a]	.55
L22		.53[a]				.72[a]	.28
L23		.66[a]				.57[a]	.43
L24		.73[a]				.47[a]	.53
L25		.77[a]				.40[a]	.60
L26		.70[a]				.50[a]	.50
L27		.73[a]				.47[a]	.53
L28		.65[a]				.58[a]	.42
L29		.51[a]				.74[a]	.26
L30		.50[a]				.75[a]	.25
L31		.60[a]				.64[a]	.36
L32		.40[a]				.84[a]	.16
L34		.34[a]				.88[a]	.12
S1 (S)			.78[b]			.39[a]	.62
S2 (S)			.78[a]			.39[a]	.62
S3 (R/L/S)	.01	.21[a]	.67[a]			.30[a]	.70
S4 (R/L/S)	-.09[a]	.22[a]	.76[a]			.25[a]	.75
S5 (L/S)		.09[a]	.80[a]			.26[a]	.74
S6 (L/S)		.14[a]	.71[a]			.33[a]	.67
W1 (W)				.93[b]		.13[a]	.87
W2 (R/L/W)	.15[a]			.64[a]		.40[a]	.60

(Table continues)

Table 23 (continued)

Item	Reading	Listening	Speaking	Writing	General	Error	SMR
			Higher-order factor structure				
					.91[a]	.17[a]	.84
					.97[a]	.07[a]	.93
					.78[a]	.39[a]	.61
					.91[a]	.17[a]	.83
			Interfactor correlations				
R	1.00						
L	.88	1.00					
S	.72	.76	1.00				
W	.83	.88	.71	1.00			
General	.91	.97	.78	.91	1.00		

Note. G = general; R = Reading; L = Listening; S = Speaking; W = Writing; SMR = squared multiple correlations.

[a] |t|> 1.96. [b] Fixed for factor scaling.

When the results of these two series of models—the versions with and without the cross-loadings of the integrated tasks—were compared, the overall goodness of model fit was roughly the same, while a considerable change in the factor loading estimates were observed for some of the Speaking and Writing items. Because the factor loadings estimated in the CFA models that allowed cross-loadings would better reflect the actual strengths of the relationships of the integrated tasks with the associated modalities, only the results based on the analyses that allowed the cross-loadings of the integrated tasks are discussed below.

In the analysis of the entire test, a series of alternative models was tested, following Rindskopf and Rose's (1988) procedure for testing relative goodness of fit of nested models, moving from testing of least restrictive to testing of more restrictive models. First, the two models described below were tested in order to establish the baseline model for the entire test:

1. Bifactor model (see Figure 4): This model hypothesized the presence of a general factor as well as four group factors corresponding to the four language modalities (Reading, Listening, Speaking, and Writing). In this model, two factor loadings were estimated for each item, one for the loading of the item on the general factor, and the other on the language modality factor with which the item was associated. For each integrated Speaking item, additional paths to Reading and/or Listening were freely

estimated as well. For the integrated Writing item, an additional path to Reading was freely estimated. The four group factors were specified as intercorrelated among themselves but uncorrelated with the general factor. This model was the least restrictive among all the models tested below.

2. Correlated trait factor model (see Figure 5): This model was nested within the bifactor model above and was obtained by trimming the general factor structure for the bifactor model. A comparison of this model with the bifactor model would indicate whether or not a global factor that directly affects all the items is present.

The fit indices for these two models are presented in the upper half of Table 24. As can be seen, the values of $\chi^2_{\text{S-B}}$/df, NNFI, CFI, and RMSEA for both models indicated good fit of these models to the data. One concern common across the two models, however, was the low GFI estimates (.82 for the bifactor Model, and .78 for the correlated trait model). In particular, the GFI estimate for the correlated trait model (.78) was fairly low. A chi-square difference test between these two models was significant at $p < .01$, suggesting that the fit of the bifactor model was significantly better than that of the correlated trait model ($p < .01$, $\chi^2_{\text{S-B difference}} = 5293.08$; *df* = 79). Moreover, the smaller ECVI for the bifactor model indicated that this model would replicate better in a different sample.

However, an inspection of the model parameter estimates suggested that the bifactor model might not be a reasonable solution. Of particular importance were (a) that most of the loadings of the items on the Reading, Listening, Speaking and Writing factors in the bifactor model were identical to those of the four trait factor loadings for the correlated trait model, and (b) that all of the loadings on the general factor in this bifactor model were nonsignificant or low. This pattern is implausible because, if a general factor had been successfully partialed out, the loadings of the items to the general factor should have been higher, while the loadings of the items to the group factors as well as the intercorrelations among the group factors should have been lower. Thus, the observed patterns in the model parameter estimates seemed to indicate a problem with model identification similar to those reported with bifactor models by previous researchers including Kunnan (1995) and Rindskopf and Rose (1988).[6] Thus, the correlated trait model was accepted as the baseline model.

Table 24

Summary of Confirmatory Factor Analysis Model Testing for the Entire Test With Additional Paths for the Integrated Tasks (All Sections)

Models	Model *df*	S-B scaled chisq	S-B scaled chisq/*df*	GFI	NNFI	CFI	RMSEA 90% CI *p*-value	ECVI 90% CI
Bifactor	2,910	5,158.48	1.77	.82	.98	.98	.017 .016–.018 1.00	2.08 2.01–2.16
Correlated traits	2,989	6,754.78	2.26	.78	.98	.98	.022 .021–.022 1.00	2.61[a] 2.52–2.70
Single trait	3,002	11,314.98	3.78	.69	.97	.97	.032 .031–.033 1.00	4.28[a] 4.16–4.40
Higher-order factor	2,991	6,855.01	2.29	.78	.98	.98	.022 .021–.022 1.00	2.65[a] 2.56–2.74

Note. S-B scaled chisq = Satorra-Bentler scaled chi-square; GFI = goodness of fit index; NNFI = non-normed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; ECVI = expected cross-validation index; CI = confidence interval.

[a] The estimated ECVI was larger than that of the saturated model.

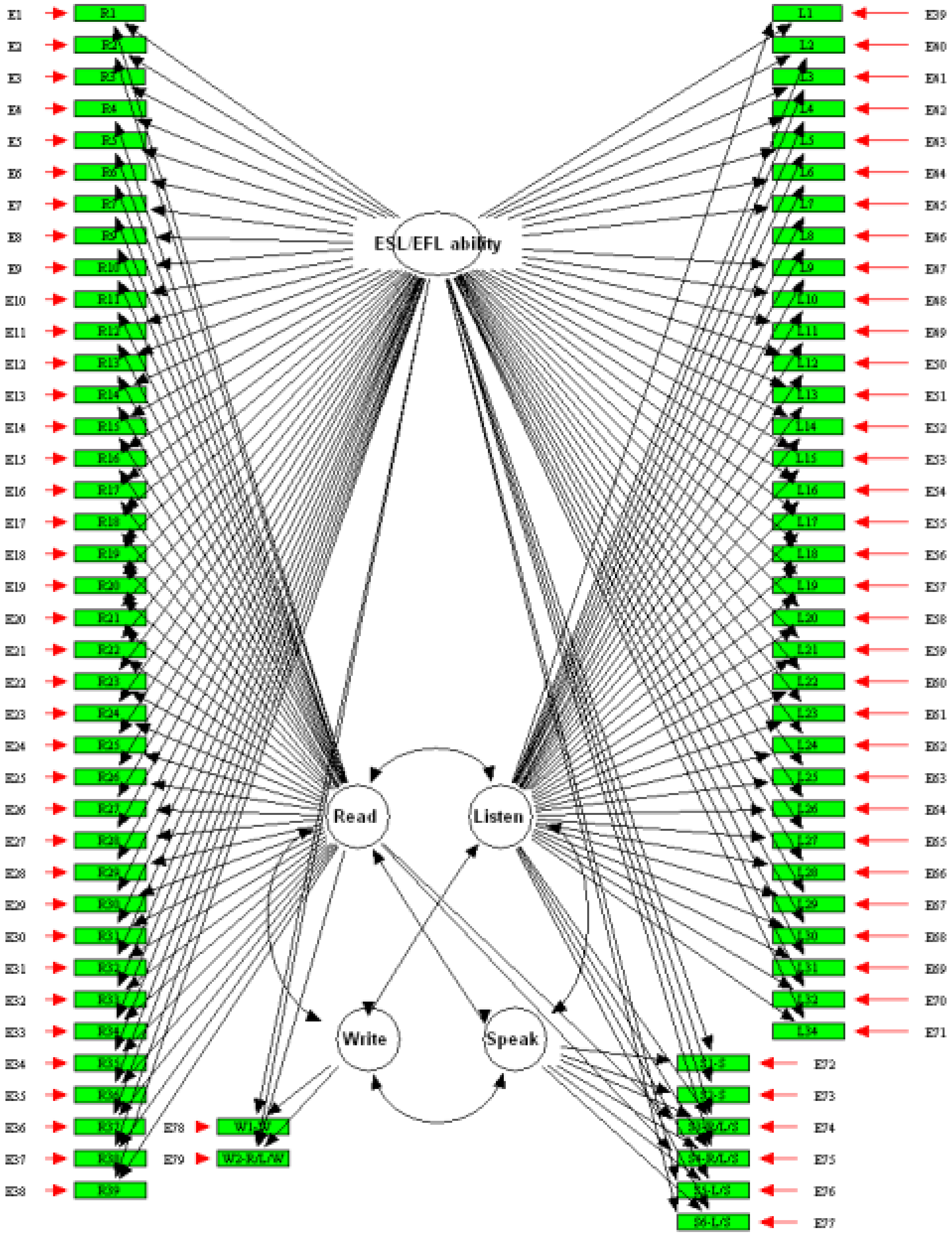


Figure 4. **Bifactor model (all sections).**

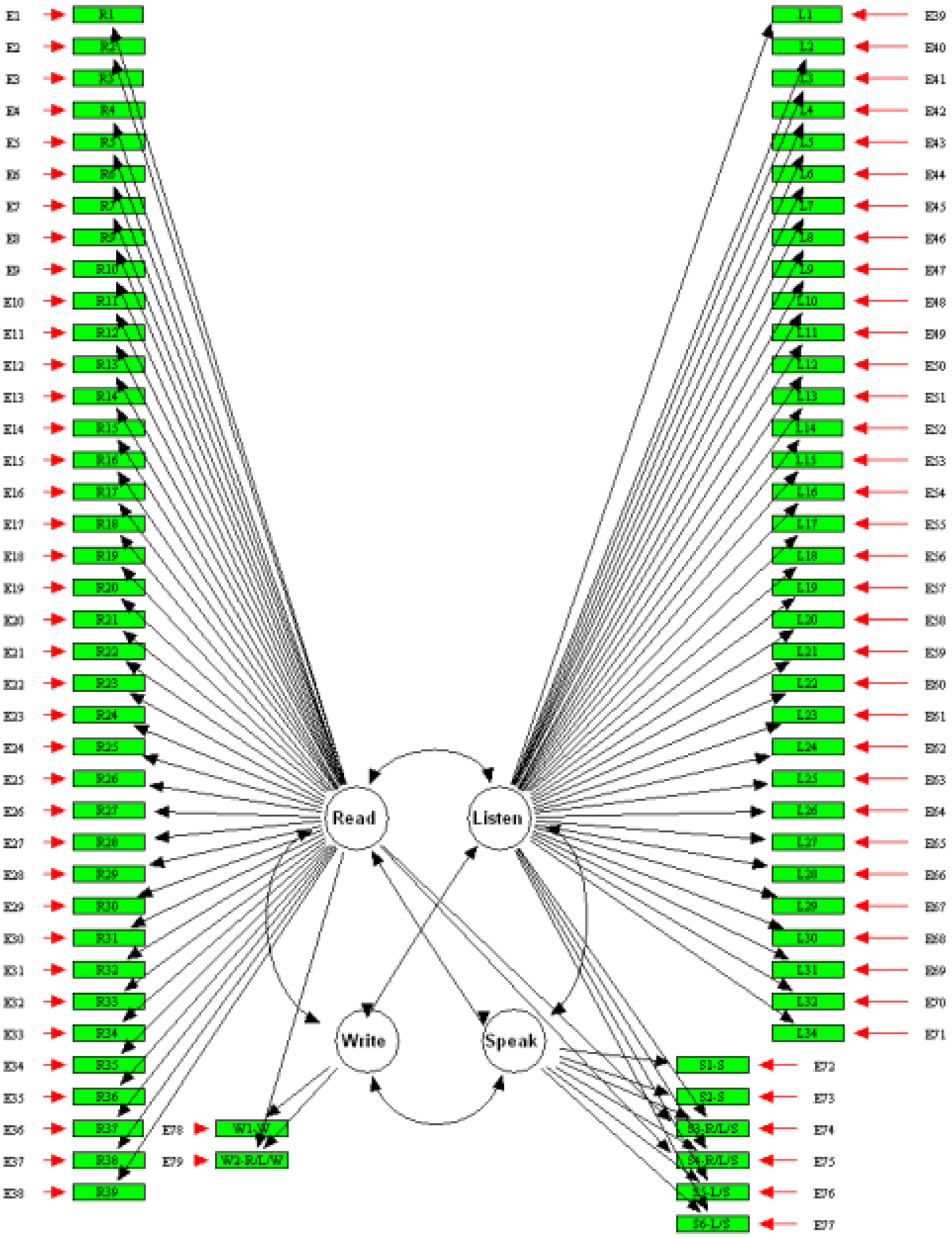


***Figure 5*. Correlated trait model (all sections).**

The trait factor structure was explored further, using Rindskopf and Rose's (1988) sequential model testing procedure. This exploration examined whether the multicomponential nature of the language abilities found in the previous language assessment literature is tenable for the TOEFL iBT as well. The two models tested below were nested within the correlated trait model:

1. Single trait model (see Figure 6): This model was obtained by fixing the interfactor correlations in the correlated trait model to 1.0. This is equivalent to saying that the four factors associated with the language modalities were indistinguishable from one another. In this model, the cross-loadings of the integrated Speaking and Writing tasks on the Reading and Listening factors would be indistinguishable from these items' loadings on the Speaking and Writing factors, because this model specifies the presence of a single trait. Thus, the cross-loadings of the integrated tasks were fixed to be the same as the loadings of the corresponding tasks on the Speaking and Writing factors. As a result, this model was identical to a model with a single factor, where only one path from the factor to each item was estimated. A comparison of this model with the correlated trait model would be a test as to whether the multicomponential nature of the language abilities assessed in the four modalities could be supported.
2. Higher-order factor model (see Figure 7): This model was obtained by imposing a constraint to the interfactor correlation structure in the correlated trait model to assume the presence of a common underlying dimension across the four modalities (i.e., the four sections of the TOEFL iBT). A conceptual distinction of this model and that of the bifactor model is that the higher-order factor model specifies the presence of a common underlying dimension that affects individual items only through the four distinct modality factors. (This differs from the specification of the general factor in the bifactor model as directly affecting the individual items.) If the higher-order factor model does not fit as well as the correlated trait model, it suggests the presence of more than one underlying dimension.

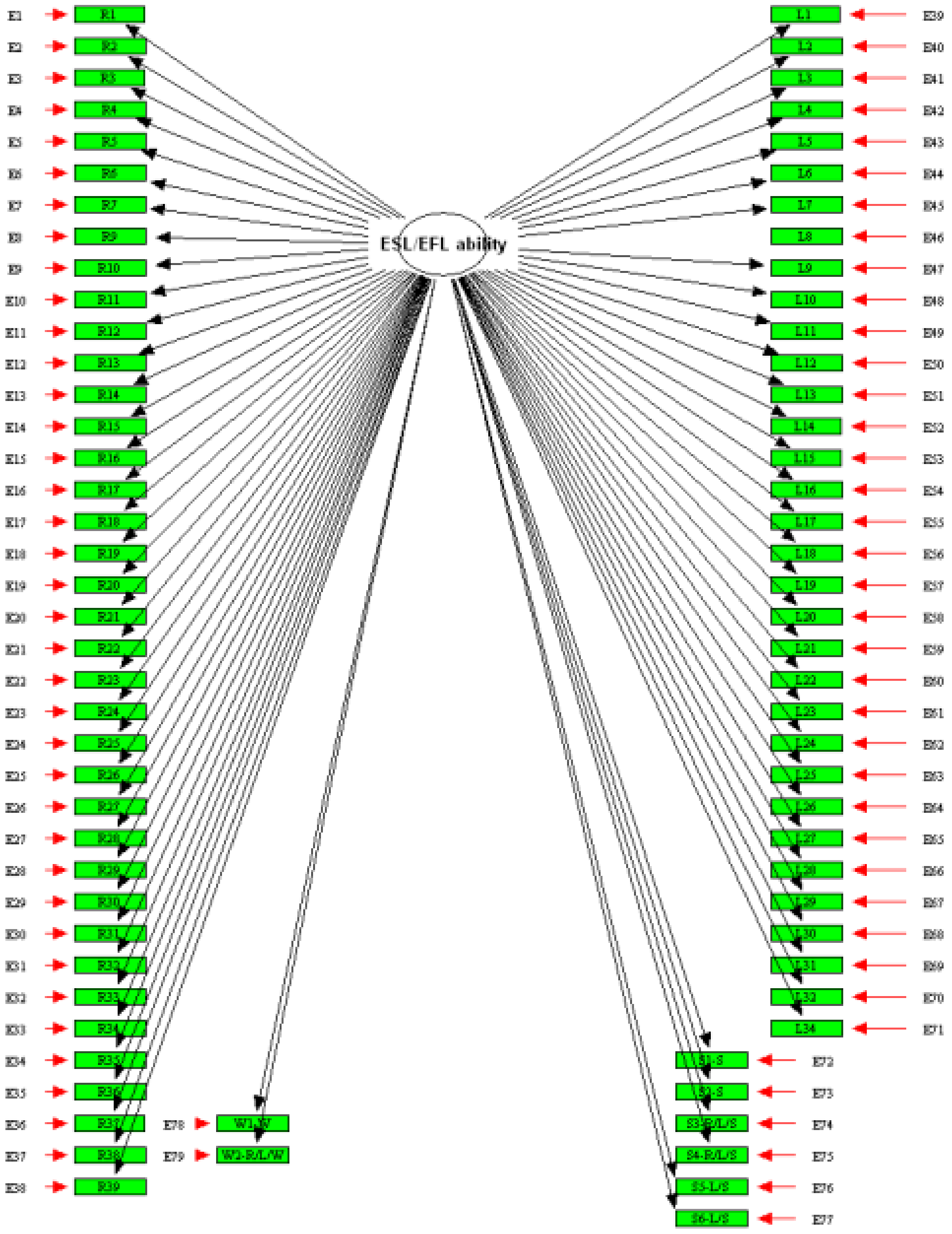


***Figure 6*. Single trait model (all sections).**

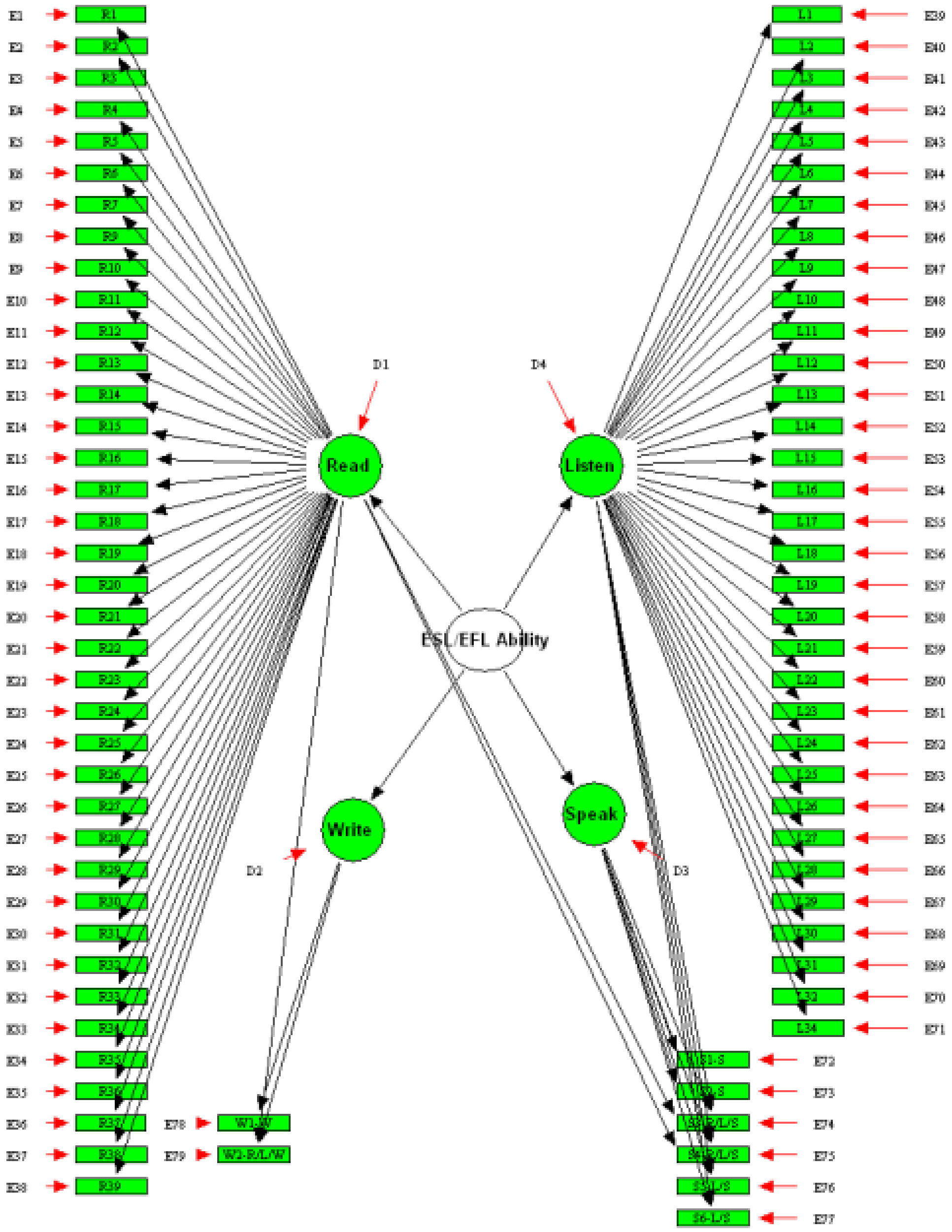


***Figure 7*. Higher-order factor model (all sections).**

The fit indices for the single trait and higher-order factor models are presented in Table 24. As can be seen, although the NNFI, CFI, and the RMSEA values for the single trait model were still respectable, the GFI (.69), $\chi^2_{\text{S-B}}$/df (3.78) and ECVI (4.28) values were considerably worse compared to those of the correlated trait model. A chi-square difference test indicated that the fit of the single trait model was significantly worse than that of the correlated trait model ($p < .01$, $\chi^2_{\text{S-B difference}} = 685.47$; $df = 13$). Thus, the specification of only one trait factor in the single trait model pointed to both a statistically and practically worse model fit compared to that of the correlated trait model.

With regard to the comparison between the correlated trait versus higher-order factor models, a chi-square difference test indicated that the fit of the correlated trait model was significantly better than that of the higher-order factor model ($p < .01$, $\chi^2_{\text{S-B difference}} = 104.91$; $df = 2$). However, the minimal differences in the $\chi^2_{\text{S-B}}$/df, NNFI, CFI, RMSEA, and ECVI values in Table 24 suggests that the fit of these two models were practically equivalent.

In summary, two conclusions can be made based on the last two model comparisons. First, the statistically and practically better fit of the correlated trait model compared to the single trait model supports the multicomponential nature of the language ability assessed in the entire TOEFL iBT test. Second, the correlated trait and the higher-order factor models indicate equivalent fit to the data. Because the higher-order factor model is more parsimonious than the single trait model, the higher-order factor model was adopted as the final model for the entire TOEFL iBT test. Additionally, the specification of the four group factors concurrently with the general factor enables a careful investigation of the relationships among the four TOEFL section scores and the total scores.

The completely standardized parameter estimates for the final model (the higher-order factor model) are presented in Table 23. With regard to the first-order factor loadings, the factor loadings of all the Speaking items on the Speaking factor and the factor loading of the Writing items on the Writing factor were consistently substantial (larger than .50), whereas the factor loadings of the Reading and Listening items on their respective factors were only moderate (less than .50 for 11 Reading items and four Listening items). The contrast between the Speaking and Writing sections versus the Reading and Listening sections is partly due to the difference in the type of data modeled. The selected-response items in the Reading and Listening sections are more susceptible to guessing by chance as well as other item characteristics that may affect examinees' response patterns than the constructed-response items in the Speaking and Writing sections. Moreover, the increased variability in the polytomously scored items in the Speaking and Writing sections (on the scales of 0-4 and 0-5, respectively) resulted in relatively large factor loadings. A similar tendency was observed among some of the polytomously scored items in the Reading section (Items R12 and R39) and one polytomously scored item in the Listening section (Item L20),[7] although, interestingly enough, the items associated with the largest factor loadings in the Reading and Listening sections were not polytomously scored items.

Another important observation concerns the substantial factor loadings of all the integrated Speaking and Writing tasks on the relevant factors (the Speaking and Writing factors, respectively). One point of interest is the magnitudes of the additional paths specified from these integrated tasks to the Reading and/or Listening factors (see Table 23). The Reading factor loadings of the integrated Speaking and Writing tasks were either not significant (Speaking Item S3) or significant but very small (Speaking Item S4 and Writing Item W2). And the Listening factor loadings of the integrated Speaking tasks were very small (ranging from .09 to .22) compared to their loadings on the Speaking factor (ranging from .67 to.80). Taken together, the pattern of the factor loadings of the integrated Speaking and Writing tasks suggest that these tasks mainly tap the target modalities.

Turning to the higher-order factor loadings in Table 23, all four sections had high loadings, ranging from .78 to .97. This supports the presence of a common underlying higher-order factor that is strongly related to the Reading, Listening, Speaking, and Writing trait factors. However, it is notable that the higher-order factor loading of the Speaking factor is somewhat lower than the loadings of the other factors, suggesting that this factor also reflects other abilities not captured by the general trait factor.

Summary. To sum up the key findings of the analysis of the entire test: (a) the higher-order factor model that included a single higher-order factor (ESL/EFL ability) and four group factors corresponding to the four modalities was a reasonable representation of the factor structure of the entire TOEFL iBT test; (b) the Speaking factor had a somewhat lower loading on the higher-order factor than did the other group factors; and (c) the patterns of the factor loadings of the integrated speaking and writing tasks suggested strong relationships of these tasks to the target modalities (i.e., Speaking and Writing, respectively).

Discussion

Internal Structure of the Reading and Listening Sections

One key finding of this study was that the three types of reading and listening abilities assessed in the Reading and Listening sections, respectively, as defined by the test specifications, were essentially unidimensional within each section. This finding is of particular interest, because some new item types in these sections were devised to assess abilities that are psychologically distinct from the abilities assessed in more conventional Reading and Listening items. One such example is the Reading to Learn items in the Reading section that are designed to assess higher-order skills that go beyond simply understanding a given text.

The finding that these new item types are not psychometrically distinct from other items in the section was similar to the results of three previous studies of the TOEFL test (Hale et al., 1988; Manning, 1987; Schedl et al., 1996). The primary interest of these studies was how new item types (multiple-choice cloze items in the Hale et al. study, cloze-elide items in the Manning study, and four different types of reasoning items in the Schedl et al. study) related to other items in the Reading section or other TOEFL sections. Schedl et al. found no evidence that the new items were psychometrically distinct from the other items in the Reading section. Hale et al. and Manning's findings agree in that the new item types studied by them had high degrees of overlap with the constructs assessed in the Structure and Written Expression and the Vocabulary and Reading Comprehension sections. These results from previous studies and the present one are consistent with Henning's (1992) simulation study results, suggesting that psychologically distinct skills and processes are not necessarily psychometrically multidimensional.

It is important to stress that the present finding that the constructs assessed in the Reading and Listening sections were essentially unidimensional does not necessarily diminish the value of distinguishing between these different types of skills and processes in the context of cognitive diagnosis. Providing fine-grained information about test performance in specific areas offers examinees guidance about issues that require more study.

Structure of the Entire Test

The second key issue was the structure of the entire test. The higher-order factor model that included a single higher-order general factor (ESL/EFL ability) and four group factors corresponding to the four TOEFL iBT sections (modalities) was the best representation of the factor structure of the entire test. This model broadly supports the reporting of five scores for the test, one for each section and a single composite score.

The higher-order factor model is consistent with the consensus in the language assessment literature that language ability is multicomponential (e.g., Bachman et al., 1995; Bachman & Palmer, 1982; Carroll, 1983; Sasaki, 1996; Shin, 2005). This hierarchical model is congruent with those found in previous confirmatory factor analyses as well (e.g., Sasaki, 1996; Shin, 2005). However, the present study yielded results that were both consistent and inconsistent with the previous factor-analytic studies of the TOEFL test discussed earlier. First, the distinct Listening factor found in the present study was consistent with the findings of Swinton and Powers (1980), Manning (1987), and two studies conducted by Hale and his associates (Hale et al., 1988, 1989) but not with the findings by Stricker et al. (2005). Second, this study identified four first-order factors corresponding to the language modality, while Stricker et al. found only two first-order factors in their analysis of the LanguEdge data. Third, this study identified a hierarchical factor structure, whereas all the other previous studies reviewed here found first-order factors only.

There are some possible explanations for the divergent findings for this study and the previous studies of the TOEFL test. The number of the distinct factors identified for this study was larger than those identified in the studies of the paper-based TOEFL test by Swinton and Powers (1980), Manning (1987) and Hale and his associates (Hale et al., 1988, 1989). This may be accounted for by difference in the content and format of the paper-based TOEFL test and the TOEFL iBT. The paper-based TOEFL test studied by Swinton and Powers (1980), Manning (1987), and Hale et al. (1988, 1989) did not include a Speaking section, and the Structure and Written Expression section in the paper-based TOEFL was not a constructed-response section where examinees were required to provide writing samples. In contrast, both the Speaking and Writing sections in the TOEFL iBT require examinees to provide speech and writing samples based on constructed-response items. Thus, the modalities and the range of language skills included in the paper-based test were narrower than those in the TOEFL iBT, which might have led to the identification of fewer factors in the paper-based test.

The difference in the range of the language skills covered in the different versions of the TOEFL test above does not explain the difference in the findings between this study and the Stricker et al. (2005) study, since the data analyzed in both studies were based on the design of the TOEFL iBT. Stricker et al. (2005) identified two correlated factors: Speaking and a fusion of Reading, Listening, and Writing.[8] In contrast, the present study identified four distinct group factors corresponding to the modalities as well as a higher-order factor.

Three issues may account for this discrepancy. First, the difference in the analytic methods used between these two studies (i.e., analysis of item parcels in the study by Stricker et al., 2005, as opposed to item-level analyses in this study) may explain the difference. However, in an unpublished investigation, the data used in the present study were analyzed by means of a confirmatory factor analysis of item parcels (Stricker & Rock, 2005). In that study a bi-factor model similar to the one tested in this study was identified. Second, Stricker et al. did not explore higher-order factor structures because of the high interfactor correlations for first-order factors .[9] This may to some extent be related to the fact that item parcels are summarized statistics, only retaining partial item information. Hence, a higher-order factor structure may have been masked by the reduced information in the variables. Nonetheless, there are also good arguments against overly specific models in relation to the typically high power of these models. Third, the nature of the samples differed in the Stricker et al. study and the present one. Stricker et al. analyzed three specific groups, Arabic, Spanish, and Chinese speakers. It is plausible that the examinees in each language group were relatively homogeneous in terms of their language development patterns and the instruction that they received. In contrast, the present study employed a combined sample of a variety of different language groups, presumably diverse in their language development and instruction.

The sample size in the present study was too small for separate analyses of language groups, given the large sample size required for the item-level confirmatory factor analysis of polychoric correlations. When large data sets for the operational TOEFL iBT become available in the future, it would be useful to investigate the factor structure within language groups.

Integrated Speaking and Writing Tasks

The factor loading patterns of the integrated speaking and writing tasks indicate that these tasks well define the target constructs and are minimally involved in the Reading and Listening constructs, the other modalities involved in the test design. This finding lays to rest the concern that the inclusion of these integrated tasks in the test might blur the interpretation of the section scores.

It is also worth noting that the somewhat weak relationship of the integrated speaking and writing tasks with the Reading and Listening tasks may be due to a design decision made for the TOEFL iBT. In the LanguEdge Courseware, an experimental TOEFL iBT prototype studied by Stricker et al. (2005), the prompts used for the integrated speaking and writing tasks were dependent on those used in the Reading and Listening sections. Examinees first completed items in the Reading and Listening sections, then they completed the integrated Speaking and Writing tasks based on the same texts they had already worked on in the Reading and Listening sections. In the current TOEFL iBT, however, the dependency of the prompt texts across the sections was removed. Thus, the reading and listening passages used for the integrated speaking and writing tasks were unique to the integrated tasks. Moreover, relatively easier reading and listening texts compared to those used in the Reading and Listening sections were employed for the integrated speaking and writing tasks, so that the difficulty of the reading and listening texts did not affect examinees' speaking and writing performances.

Limitations

Some limitations of the present study should be noted. First, the sample size was only marginally acceptable for the item-level CFA of 79 items included in the entire test, and was too small for separate analyses of subgroups, such as native language groups and ability groups (Jöreskog & Sörbom, 1996, p. 171). Second, this study used data collected in a field study of the TOEFL iBT, and there may be differences between that sample and the population that will be taking the TOEFL iBT in the future, in terms of language groups, countries, ability level, test-taking motivation, and familiarity with the test. Third, the model fit was far from ideal. Fourth, a full investigation of the structure of the Writing section was not possible because the design of this section incorporated only two items. For all of these reasons, the results of this study should be interpreted with caution, and a replication should be conducted, for different groups, with examinees taking the operational TOEFL iBT.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1-34.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language*: *The Cambridge-TOEFL comparability study.* Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning, 31,* 67–86.

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

Bagozzi, R. P., & Yi, Y. (1992). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*, 74–94.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. MS-21). Princeton, NJ: ETS.

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

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series No. MS-20). Princeton, NJ: ETS.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carr, N. T. (2003). *An investigation into the structure of text characteristics and reader abilities in a test of second language reading*. Unpublished doctoral dissertation, University of California, Los Angeles.

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

Chen, F. F., West, S. G., & Sousa, K. H. (2006). Comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189-225.

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

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series No. MS-18). Princeton, NJ: ETS.

Davidson, F. D. (1988). *An exploratory modeling of the trait structures of some existing language test datasets.* Unpublished doctoral dissertation, University of California, Los Angeles.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. MS-17). Princeton, NJ: ETS.

ETS. (2003). *TOEFL test and score data summary: 2002-2003 test year data*. Princeton, NJ: Author.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 11–28). Rowley, MA: Newbury House.

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. 32). 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. 26). Princeton, NJ: ETS.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing, 9,* 1–11.

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 Publications.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary* *developments in mathematical psychology: Measurement, psycho-physics, and neural information processing* (Vol. 2, pp. 1–56). San Francisco: Freeman.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide.* Chicago, IL: Science Software International.

Jöreskog, K., & Sörbom, D. (2003a). PRELIS (Version 2.54) [Computer software]. Chicago, IL: Scientific Software International.

Jöreskog, K., & Sörbom, D. (2003b). LISREL (Version 8.54) [Computer software]. Chicago, IL: Scientific Software International.

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

Kunnan, A. (1995). *Test taker characteristics and test performance: A structural equation 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-87-18). Princeton, NJ: ETS.

Marsh, H. W. (1988). Multitrait-multimethod analysis. In J. P. Keeves (Ed.), *Educational research methodology, measurement, and evaluation: An international handbook* (pp. 570–580). Oxford, England: Pergamon.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*(1), 47–70.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 177–198). Thousand Oaks, CA: Sage.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum.

Nissan, S., & Sawaki, Y. (2005, March). *Can TOEFL help with classroom harmony?* Paper presented at the 39th TESOL convention and exhibit, San Antonio, Texas.

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

Oller, J. W. (1983). A consensus for the eighties? In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 351–356). Rowley, MA: Newbury House.

Oller, J. W., & Hinofotis, F. A. (1980). Two mutually exclusive hypotheses about second language ability: Factor analytic studies of a variety of language subtests. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing* (pp. 13–23). Rowley, MA: Newbury House.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Erlbaum.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York: Lang.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity, 24*, 367–386.

Satorra, A., & Bentler, P. (1999). *A scaled difference chi-square test statistic for moment structure analysis* (UCLA Statistics Series No. 260). Los Angeles: University of California.

Sawaki, Y., & Lee, Y-.W. (2006, April). *A comparison of three cognitive diagnosis models in modeling examinees' reading and listening skill mastery patterns.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schedl, M. A., Gordon, P. C., & Tang, K. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items* (TOEFL Research Rep. No. 53). Princeton, NJ: ETS.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

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. (2005). *Factor analysis of New Generation TOEFL.* Unpublished manuscript.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series No. MS-32). Princeton, NJ: ETS.

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. 6). Princeton, NJ: ETS.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.

Zhang, Y., DiBello, L., Puhan, G., Henson, R., & Templin, J. (2006, April). *Estimating skills classification reliability of student profiles scores for TOEFL Internet-based test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Notes

[1] In the dataset used for the present study, half points on the individual Writing item scores were rounded up to the closest integer.

[2] The test specifications of the Speaking and Writing sections were also derived from the corresponding framework papers. See Butler, Eignor, Jones, McNamara, and Suomi (2000) and Cumming, Kantor, Powers, Santos, and Taylor (2000) for details.

[3] One Basic Comprehension item that was not scored is excluded from the number.

[4] One Pragmatic Understanding item that was not scored is excluded from the number.

[5] The item types within each of these reading purposes were not modeled in the MTMM analysis because too few items represented some of the item types.

[6] See Chen, West, and Sousa (2006), however, for an example where a bifactor model was adopted over a higher-order factor model in the personality domain.

[7] Item R26 in the Reading section and Item L4 in the Listening section were polytomously scored items as well. However, the factor loadings of these items were not particularly high compared to those of other items.

[8] The correlated two-factor model supported by Stricker et al. (2005) was also tested as a supplementary analysis in this study. This model resulted in a proper and interpretable solution with the interfactor correlation of .79. However, it was not pursued because the model chi-square value ($\chi^2_{\text{S-B}}$ = 8,684.64; *df* = 3,001) and its ratio to the model degrees of freedom ($\chi^2_{\text{S-B}}/df$ = 3.55) as well as the model fit indices (GFI = .74, NNFI = .97, CFI = .97, RMSEA = .026, ECVI = 3.31) suggested that the fit of this model is relatively poor compared to that of the correlated trait factor and higher-order factor models (cf. Table 24).

[9] Stricker et al. (2005) mentioned their original plan was to test a higher-order factor model if a correlated four-factor model, similar to the baseline model for the entire test initially adopted in this study, fit the data.

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