DOCUMENT RESUME

ED 395 029 TM 025 047

AUTHOR	Hale, Gordon A.; And Others
TITLE	Confirmatory Factor Analysis of the Test of English as a Foreign Language. TOEFL Research Reports, Report 32.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-42
PUB DATE	Dec 89
NOTE	65p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Ability; Adults; *English (Second Language); *Factor Structure; *Language Tests; *Listening Comprehension; *Research Methodology; Test Construction
IDENTIFIERS	*Confirmatory Factor Analysis; *Test of English as a Foreign Language

ABSTRACT

Previous studies found inconsistent results about the factor structure of the Test of English as a Foreign Language (TOEFL), with one study finding a two-factor structure and the other, a three-factor solution. This study investigated those inconsistencies and provided further information about the TOEFL factor structure. It was hypothesized that the inconsistency between studies was related to the populations under investigation, as the earlier study used TOEFL examinees in both domestic and overseas test centers, whereas the more recent study used domestic examinees only. The present data did not support this hypothesis, however. Confirmatory factor analyses were conducted for each of several language groups, using data from a 1984 TOEFL. These analyses yielded essentially similar results for domestic and overseas populations, as well as for the combined population. In all cases, the data supported a two-factor interpretation, with the two factors related to the Listening Comprehension section and the nonlistening sections. Additional study indicated that the use of different factor analytic methodology in the two previous studies undoubtedly contributed to the inconsistency, although further study would be needed to determine exactly what aspects of the methodology played a role. Findings also indicated that the basic factor structure of the test did not change substantially between 1976 and 1984 and that a two-factor structure was in evidence for examinees in high- and low-proficiency groups. Appendixes give the factor loadings for the two-factor solution and correlations between factors in two-factor confirmatory factor analyses. (Contains 15 tables and 22 references.) (SLD)

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 32
DECEMBER 1989

Confirmatory Factor Analysis of the Test of English as a Foreign Language

Gordon A. Hale
Donald A. Rock
Thomas Jirele

ETS®
EDUCATIONAL TESTING SERVICE

Confirmatory Factor Analysis of the Test of English as a Foreign Language

by

Gordon A. Hale, Donald A. Rock, and Thomas Jirele

Educational Testing Service
Princeton, New Jersey 08541

RR-89-42

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Acknowledgments

The authors express their sincere appreciation to Marian Helms and Ruth Yoder for secretarial assistance, to Irene Kostin for help in data analysis, to Grant Henning and Warren Willingham for general advice, and to Robert Boldt, Philip Oltman, Peter Pashley, and Lawrence Stricker for reviewing an earlier draft of this manuscript.

The authors also gratefully acknowledge the advice of the following persons at various stages of the preparation and conduct of this study: Lyle Bachman, University of Illinois; Charles Stansfield, Center for Applied Linguistics; and John H. A. L. de Jong, National Institute for Educational Measurement (Cito), Arnhem, Netherlands.

Finally, the authors wish to thank the TOEFL Research Committee for its support of the project.

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were 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, and 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.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide the data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1989-90) members of the TOEFL Research Committee are:

Patricia L. Carrell (Chair)	University of Akron
Lily Wong Fillmore	University of California at Berkeley
Fred Genesee	McGill University
Frederick L. Jenks	Florida State University
Elliott Judd	University of Illinois at Chicago
Elizabeth C. Traugott	Stanford University

Abstract

In a recent study, confirmatory factor analyses indicated that, for each of several language groups, TOEFL performance can be characterized by two factors, associated with (a) the Listening Comprehension section and (b) the other sections of the test (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1988). This conclusion was inconsistent with that drawn in the most comprehensive previous factor-analytic study of the TOEFL, which suggested a three-factor solution for each of several languages (Swinton & Powers, 1980). The present study investigated the inconsistency in conclusions drawn from these two studies and provided further information about the factor structure of the TOEFL.

It was hypothesized that the inconsistency between studies was related to the populations under investigation, as the earlier study used TOEFL examinees in both domestic and overseas test centers, whereas the more recent study used domestic examinees only. The present data did not support this hypothesis, however. Confirmatory factor analyses were conducted for each of several language groups, using data from the test form used in the more recent study (a 1984 TOEFL), and these analyses yielded essentially similar results for both domestic and overseas populations, as well as for the combined population. In all cases, the data supported a two-factor interpretation, with the two factors related to the Listening Comprehension section and to the nonlistening sections.

Other hypotheses were that the inconsistency between studies was due to differences in factor-analytic methodologies used or to changes in the test over time. Confirmatory factor analyses were conducted using the data from the earlier study (taken from a 1976 TOEFL) as well as data from the more recent study. These analyses supported a two-factor interpretation in both cases, for each of several language groups, with the two factors associated with Listening Comprehension and with the other sections of the test. Thus, the use of different factor-analytic methodologies in the two previous studies undoubtedly contributed to the inconsistency, although further work would be needed to determine exactly what aspects of the methodology played a role. The data also provided tentative evidence that the basic factor structure of the test may not have changed substantially between 1976 and 1984.

The role of examinee proficiency in determining the TOEFL's factor structure was also examined. For each of several language groups, low- and high-proficiency groups were defined on the basis of approximately one third of the items in the TOEFL, drawn from all sections of the test, and factor analyses were then performed on an abbreviated TOEFL, which consisted of the remaining items. For both proficiency levels within each language group a two-factor structure appeared to underlie performance, with the factors once again linked to Listening Comprehension and to the other, nonlistening sections of the test.

Table of Contents

Page

Introduction 1

Hypotheses Regarding the Inconsistency Between Studies 1
Factor Structure of the TOEFL for Low- and for High-Proficiency Examinees 4

Method 7

Subjects 7
The TOEFL 7
Methods of Analysis 8

Results 11

Comparison of Domestic and Overseas Examinees 11
Comparison of November 1976 and November 1984 Tests 18
Comparison of Low- and High-Proficiency Examinees 25

Discussion 33

Comparison of Domestic and Overseas Examinees 33
Comparison of November 1976 and November 1984 Tests 33
Comparison of Low- and High-Proficiency Examinees 34
Examination of Other Factor-Analytic Studies Involving the TOEFL . . 35
Conclusions and Practical Implications 36

References 39

Appendices 41

A. Factor Loadings for Two-Factor Solution 41

B. Correlations Between Factors in Two-Factor Confirmatory Factor Analyses 49

List of Tables

Table		Page
1	Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the November 1984 TOEFL for Each Language Group, Tested in Domestic and Overseas Centers	12
2a	Domestic Test Centers: Correlations Among TOEFL Subscores and Subscore Reliabilities for Each Language Group (November 1984 TOEFL)	13
2b	Overseas Test Centers: Correlations Among TOEFL Subscores and Subscore Reliabilities for Each Language Group (November 1984 TOEFL)	14
3a	Domestic Test Centers: Indices from Confirmatory Factor Analyses of the TOEFL for One-, Two-, and Three-Factor Solutions (November 1984 TOEFL)	16
3b	Overseas Test Centers: Indices from Confirmatory Factor Analyses of the TOEFL for One-, Two-, and Three-Factor Solutions (November 1984 TOEFL)	17
4	Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the TOEFL for Each Language Group, for November 1976 and November 1984 Tests (Domestic and Overseas Centers Combined)	19
5a	November 1976 TOEFL: Correlations Among Subscores and Subscore Reliabilities for Each Language Group (Combined Domestic and Overseas Populations)	21
5b	November 1984 TOEFL: Correlations Among Subscores and Subscore Reliabilities for Each Language Group (Combined Domestic and Overseas Populations)	22
6a	November 1976 TOEFL: Indices from Confirmatory Factor Analyses for One-, Two-, and Three-Factor Solutions (Combined Domestic and Overseas Populations	23
6b	November 1984 TOEFL: Indices from Confirmatory Factor Analyses for One-, Two-, and Three-Factor Solutions (Combined Domestic and Overseas Populations)	24
7	Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the Abbreviated 1984 ıOEFL for Each Language Group, for Low- and High-Proficiency Examinees (Combined Domestic and Overseas Populations)	26

List of Tables (Continued)

Table | Page

8a Low-Proficiency Examinees: Correlations Among Subscores and Subscore Reliabilities for Each Language Group on the Abbreviated 1984 TOEFL (Combined Domestic and Overseas Populations) 28

8b High-Proficiency Examinees: Correlations Among Subscores and Subscore Reliabilities for Each Language Group on the Abbreviated 1984 TOEFL (Combined Domestic and Overseas Populations) 29

9a Low-Proficiency Examinees: Indices from Confirmatory Factor Analyses of the Abbreviated 1984 TOEFL for One-, Two-, and Three-Factor Solutions (Combined Domestic and Overseas Populations) 30

9b High-Proficiency Examinees: Indices from Confirmatory Factor Analyses of the Abbreviated 1984 TOEFL for One-, Two-, and Three-Factor Solutions (Combined Domestic and Overseas Populations) 31

Introduction

In a recently completed study (Hale, Stansfield, Rock, Hicks, Butler & Oller, 1988), confirmatory factor analyses of the TOEFL for each of nine major language groups indicated that two factors can account for the variance in TOEFL scores. One factor was related to the Listening Comprehension section and the other, to the nonlistening sections (Structure and Written Expression, Vocabulary and Reading Comprehension).

These results contrasted with those obtained in the most comprehensive previous factor-analytic study of the TOEFL (Swinton & Powers, 1980), which looked at the factor structure of the test for each of seven major language groups. Data from that study suggested that three factors underlie TOEFL performance for most language groups, albeit with some differences among language groups in definition of the three factors.

The failure of Hale et al. to observe a three-factor structure for the TOEFL was unexpected, given the assumption--based partly on the Swinton and Powers study--that three separate aspects of English proficiency are tapped by the TOEFL. The present research was conducted, therefore, as a follow-up to the study by Hale et al., in an effort to account for the inconsistency in implications of these two studies, and to provide further information about the factor structure of the TOEFL.

Although there are a few other published factor-analytic reports that have involved the TOEFL, which will be considered in the Discussion section, the studies by Hale et al. and Swinton and Powers served as the basis for the present study for several reasons. First, these studies provided the most comprehensive factor analyses of the TOEFL, as they were based on large numbers of examinees taking operational administrations of the test; other studies have generally been limited in scope, employing fewer than the several hundred subjects required for a proper factor analysis. Second, these two studies included separate analyses for each of several major language groups; other studies have usually been done with a single language group, which limits generalizability, or with combined language groups, which adds variance associated with language group differences. Third, both of these studies used the current three-part TOEFL, whereas most other factor-analytic studies of the TOEFL employed the five-part version of the test that was in use prior to 1976 (see Discussion).

Hypotheses Regarding the Inconsistency Between Studies

The following hypotheses were advanced to account for the inconsistency in conclusions suggested in the studies by Hale et al. and Swinton and Powers.

The first hypothesis: Domestic versus overseas populations. One of the major differences between these two studies lay in the population of examinees used. Hale et al. employed only domestic examinees--that is, foreign students tested in centers in the United States and Canada. In the study by Swinton and Powers, on the other hand, a sample was drawn from the combined population of overseas and domestic examinees. It was hypothesized that, when one examines data for domestic examinees only, the apparent factor structure of the test will differ from that observed when data from overseas examinees, or from the combined population of domestic and overseas examinees, are used.

A theoretical basis for this hypothesis (from Anastasi, 1970) is that exposure to a relatively standardized curriculum across schools should lead to the appearance of a broad factor, whereas a curriculum that varies from school to school should lead to greater differentiation among factors. TOEFL examinees in domestic test centers are typically enrolled in intensive English courses that tend to be similar to each other, in that they provide a balanced emphasis on teaching of such aspects of English proficiency as grammar, reading, and vocabulary. As a result, the same students who are relatively proficient in one area are relatively proficient in the others as well, thus leading to a greater similarity among factors than would be expected with a less standard curriculum. For overseas examinees, however, it is believed that there is less uniformity among the situations in which English is learned and, thus, a lower probability that skills in various areas will improve in a uniform manner. Consequently, factor analysis might be expected to show greater differentiation among factors for overseas examinees.

There was a practical basis for testing this hypothesis as well. Research on the TOEFL occasionally is conducted with domestic examinees only. An assumption is that domestic and overseas populations are similar, and indeed, statistical analyses conducted thus far have shown a reasonably high degree of similarity between these populations with respect to means and variances of scores (albeit with some variation across test sections), item-total score correlations, and other statistics. It has remained to be determined, however, whether the basic factor structure of the test is the same for these two subsets of the examinee population.

To address the above hypothesis, confirmatory factor analyses were performed separately for overseas examinees and domestic examinees, using the November 1984 test studied by Hale et al. (The analysis for domestic examinees essentially constituted a reanalysis of the Hale et al. data, although with slight changes in the data sets as noted below). Analyses were also performed for the combined population of domestic and overseas examinees. Separate analyses were conducted for each of five language groups, Arabic, Chinese, Farsi, Japanese, and Spanish, the five language groups common to both the studies by Hale et al. and by Swinton and Powers.

A second hypothesis: The statistical methods used. Differences in analytic methods could also account for the apparent inconsistency between the two studies cited above. Hale et al. used confirmatory factor analysis, employing current LISREL procedures (Joreskog & Sorbom, 1981, 1983), whereas

Swinton and Powers used other methods.[1] Also, in the study by Swinton and Powers, individual items were the units of analysis, and factor analyses were performed on tetrachoric correlations among individual items. In the study by Hale et al., on the other hand, the units of analysis were parcels of items--that is, sets of items within each subsection of the test that were intended as replicates of each other. It has been noted that use of individual items can sometimes produce a factor associated with item difficulty (Hulin, Drasgow, & Parsons, 1983). By contrast, the parcels used by Hale et al. were selected randomly within each item subtype, with the constraint that all parcels would have similar overall difficulty levels, thereby eliminating the possibility that a factor associated with item difficulty would appear.

To address the above hypothesis, the data used by Swinton and Powers (taken from the November 1976 administration of the TOEFL) were analyzed using the methods employed by Hale et al. That is, confirmatory factor analysis was employed, using LISREL procedures, and item parcels served as the units of analysis. If the results proved to be similar to those observed for the November 1984 test, it could be inferred that differences in factor-analytic methodologies likely played a role. In this case, it would have to be left for further research to determine exactly which aspects of analytic procedure may have contributed to the apparent difference in factor structures suggested by the two previous studies. The data would at least show, using current methods of confirmatory factor analysis, the extent to which the three sections of the TOEFL can be regarded as measuring distinct aspects of proficiency.

A third hypothesis: Changes in the test over time. It is also possible that the difference in implications of the studies by Hale et al. and by Swinton and Powers was due to a change in the test over time. Even though the basic three-section structure of the test remained invariant between 1976 and 1984, it is possible that subtle changes could have occurred in implementation of the test specifications, such that the nonlistening sections of the test--Structure and Written Expression, and Vocabulary and Reading Comprehension--gradually came to measure less distinct aspects of proficiency.

[1] Swinton and Powers first obtained a Varimax solution and then subjected a four-factor solution to an orthogonal Procrustean rotation to force the data to fit the design structure of the test. In the latter analysis, one target factor was specified as maximum loadings on selected Listening Comprehension items; a second target factor, maximum loadings on selected Structure items; and a third target factor, maximum loadings on selected Vocabulary items. Thus, the analysis was confirmatory in intent but differed from the LISREL confirmatory procedures used here in certain respects, such as (a) the factors were forced to be orthogonal to each other and (b) the second and third factors were defined in relation to items in only one of the two parts of TOEFL Section 2 (Structure) and one of the two parts of Section 3 (Vocabulary).

Comparison of factor-analytic results for the November 1976 and November 1984 tests, using the same methods of analysis in each case, would help address this hypothesis. If the factor-analytic results were found to be essentially the same for both tests, it might be inferred that the factor structure of the test had not changed substantially over that time period. (Note, however, that the 1976 and 1984 test forms were administered to different cohorts--that is, populations differing in time--so that firm conclusions about the comparability of 1976 and 1984 test forms could not be drawn unless the two forms were administered to samples taken from the same cohort.)

Factor Structure of the TOEFL for Low- and for High-Proficiency Examinees

An additional objective of the study was to determine whether the English proficiency of the examinees plays a role in determining the factor structure of the TOEFL. Two different theoretical positions lead to opposite predictions. One hypothesis suggests decreasing factor differentiation as a function of increasing proficiency. Higgs and Clifford (1982) have argued that, for interview performance at least, second language acquisition includes a progression toward more equal involvement of the various skills involved in performance. Similarly, studies conducted within an interlanguage framework (cf. Selinker, 1969) suggest that it is reasonable to hypothesize a decreasing differentiation of factors in second language performance with increasing acquisition of the second language. This view derives from the notion that the effect of transfer from the first language is believed to be greatest at the beginning of acquisition of the second language, then tends to diminish as proficiency in the second language increases.

Another hypothesis, however, maintains that the degree of differenriation among skills increases with d.velopment of those skills (cf. discussion by Anastasi, 1970). In the early stages of learning, a broad factor is apparent, due to the fact that some learners are of low proficiency in many skills, while others are of higher proficiency in those skills. As learning progresses, howevever, greater differentiation among skills becomes apparent. This is partly due to the fact that some learners begin to excel in certain skills, while other learners begin to excel in others. This is especially true if learners are exposed to a curriculum that becomes increasingly differentiated into separate areas, with different skills taught by different teachers in different classrooms.

A recent study by Oltman, Stricker, and Barrows (1988), using multidimensional scaling and cluster analysis, obtained results consistent with the first of these two hypothesis. When those authors examined the dimensional structure of the test separately for low-, medium-, and high-proficiency examinees within each of several language groups, they found a clearer differentiation among dimensions for low-proficiency examinees than for those of medium- or high-proficiency.

At the same time, there is at least indirect evidence for the second hypothesis. Swinton and Powers (1980) found that, of seven language groups studied, the group that was substantially lower in proficiency than the others--the Farsi speakers--was the one group for whom there was considerably less differentiation in factor structure of the TOEFL. Also, Oller and Hinofotis (1980) studied Iranian students (Farsi speakers) and found that, after a general factor was partialed out, there was no remaining variance to be explained in TOEFL performance; however, in another portion of their study involving a mixed group of foreign students, separate factors emerged that were associated with TOEFL Listening Comprehension and with the other sections of the test.

To examine the relation of examinee proficiency to differentiation among factors in the TOEFL, the present study included factor analyses for low- and high-proficiency examinees, separately for each of several language groups. Proficiency was defined by performance on a subset of approximately one third of the items on the TOEFL, selected from all sections, and factor analyses were performed on data from the remaining items. In this way it was possible to achieve independence of the test used to define proficiency and the test used in the factor analysis, thus providing a more appropriate test than would be the case if data from the overall TOEFL were used for both purposes.

Method

Subjects

The subjects of the study were 14,974 examinees who took the TOEFL at the November 1984 administration, and 4,659 examinees who took the TOEFL at the November 1976 administration. Included were both "domestic" examinees, who took the TOEFL in the United States or Canada, and "overseas" examinees, who took the TOEFL in other countries.

The sample consisted of examinees from the five language groups that were common to both the Hale et al. (1988) and Swinton and Powers (1980) studies: Arabic, Chinese, Farsi, Japanese, and Spanish. Excluded from the group of Chinese speakers were (a) students from the People's Republic of China, since this subgroup is believed to differ in important respects from other Chinese-speaking groups and was essentially absent from the Swinton and Powers sample, and (b) students from Taiwan, since Taiwanese students were excluded from the Swinton and Powers sample for procedural reasons. With these exceptions, data were analyzed for all examinees in the five language groups taking the November 1984 test, and for the random sample of approximately 1,000 examinees in each language group taking the November 1976 test that were studied by Swinton and Powers. (Numbers of subjects per language group for each test are indicated in Table 4 in the Results section.) The sample thus contained representatives of four different language families: Indo-European (Spanish and Farsi), Altaic (Japanese), Sino-Tibetan (Chinese), and Semitic (Arabic).

The TOEFL

The three sections of the TOEFL are (a) Listening Comprehension, (b) Structure and Written Expression, and (c) Vocabulary and Reading Comprehension (Educational Testing Service, 1987). The item types are as follows. In Listening Comprehension, the examinee hears spoken material (either single statements, short dialogues, or short monologues) and then hears questions about them, which he or she answers by selecting the correct answer choices in a test booklet. In the other two sections, all information is presented in written form. The Structure and Written Expression section consists of two item subtypes. In each Structure item, the examinee is given a sentence from which a phrase has been deleted and, from the response alternatives, must choose the word or phrase that best fits into the sentence. In each Written Expression item, the examinee is given a sentence in which words or phrases are underlined, and the examinee must indicate the underlined word or phrase that is ungrammatical. The third section consists of two basic item subtypes, Vocabulary and Reading Comprehension. In each Vocabulary item, a sentence is presented with a word or phrase underlined, and the examinee must choose the response alternative that is synonymous with the underlined word or phrase. In the Reading Comprehension part, segments of text are presented and, following each segment, several questions appear. The examinee must answer each question by selecting the best response alternative.

Methods of Analysis

Each of the five subsections of the test mentioned above was divided into "parcels," with each parcel consisting of the total score for a group of items. There were three parcels of items for each subsection except Structure, for which there were two parcels. The items in each parcel were chosen so that, within each subsection, the parcels would be roughly equal in average difficulty and distribution of item difficulties. These parcels served as the basic units of analysis. The numbers of items in the various parcels are shown below. (Note that, while there are 150 total items in the TOEFL, four items are nonoperational, and only the 146 operational items are included in the present analyses.)

Description of Parcels

Parcel No.	Label	No. of Items
1	Listening Comprehension 1 (LC1)	17
2	Listening Comprehension 2 (LC2)	17*
3	Listening Comprehension 3 (LC3)	16
4	Structure 1 (S1)	7*
5	Structure 2 (S2)	7
6	Written Expression 1 (WE1)	8*
7	Written Expression 2 (WE2)	8
8	Written Expression 3 (WE3)	8
9	Vocabulary 1 (V1)	10
10	Vocabulary 2 (V2)	10
11	Vocabulary 3 (V3)	9*
12	Reading Comprehension 1 (RC1)	10*
13	Reading Comprehension 2 (RC2)	10
14	Reading Copmrehension 3 (RC3)	9

* Parcels that were removed from the "abbreviated TOEFL" and used to determine proficiency level in analyses comparing low- and high-proficiency examinees.

Thus, there were three parcels for each of the test's subsections except Structure, for which there were two parcels. It was determined that, to permit analyses of up to five factors (one for each subsection), each factor should be represented by two, and preferably three, parcels in order to have a sufficient basis for definition of each factor. It was also determined that there should be at least seven items per parcel to provide relatively stable measurement per parcel (hence the decision to employ only two parcels in the Structure section). It should be noted that the differences between test sections in numbers of items per parcel have no bearing on the factor pattern to emerge from the analyses. While parcel size would relate to the magnitude of the factor loadings, it would have no effect on the comparisons of interest in the study.

In most of the factor analyses conducted here, variance-covariance matrices based on the parcel scores were derived, and these matrices provided the input to the factor analyses. For reasons discussed below, factor analyses involving the low- and high-proficiency groups were based on correlations among parcel scores rather than variance-covariance matrices.

The mode of analysis used in this study was confirmatory factor analysis following the maximum likelihood estimation procedures of LISREL VI (Joreskog & Sorbom, 1981, 1983). No exploratory analyses were performed. (Note that Hale et al. [1988] performed initial exploratory analyses before confirmatory factor analyses. In each of the principal components analyses they reported, the first eigenvalue was relatively large, but the second eigenvalue was close enough to 1 to suggest the value of investigating at least a two-factor solution as well as a one-factor solution.) While it would have been possible first to perform a principal components analysis here (followed, for example, by an oblique rotation), it was determined that the issues under investigation could best be addressed by going directly to confirmatory analyses. With a confirmatory procedure, factors are defined a priori, in this case according to the test content, and the analyses determine the extent to which the empirically determined factor structure of the test corresponds to key content distinctions.

Three models, involving one, two, and three factors, provided the major bases for comparisons in this study. The single factor in the first model was defined in terms of performance on the entire test. The two factors in the second model were defined as (a) Listening Comprehension and (b) the other sections of the test. The three factors in the third model were the three sections of the test: (a) Listening Comprehension, (b) Structure and Written Expression, and (c) Vocabulary and Reading Comprehension. In addition, a null model was tested, in which no common factors were presumed to underlie the data, primarily to provide data needed for computation of indices for the three models under study. Also, a five-factor model, involving all five subsections of the test, was tried in order to provide additional data in case none of the three models indicated above sufficiently fit the data. Data from the five-factor model are not presented here, however, given that the simpler models provided a satisfactory fit, as discussed below.

Four goodness of fit indicators were derived from each analysis, as described below. Each of these indices provides a measure of the degree to which a particular model fits the data.

Goodness of fit index (GFI). The GFI, which ranges from 0 to 1.00, indicates the relative amount of variance and covariance jointly accounted for by the factor model (Joreskog & Sorbom, 1985). The GFI was originally thought to be independent of sample size, but recent simulation studies suggest that this may not be the case (Marsh, personal correspondence, September 1987). In the present case, however, relevant comparisons are between one-, two- and three-factor models within populations, so sample-size dependence is irrelevant here.

Root mean square residual (RMSR). This is the average covariance among parcels that is left over after the hypothesized model has been fitted (Joreskog & Sorbom, 1985). The RMSR is estimated independently of sample size. When the solution is based on the variance/covariance matrices, the RMSR is interpreted with respect to the size of the original matrix of covariances. Dividing the RMSRs by the root mean squares indicates the percentages of variance not accounted for by the hypothesized model.

Tucker-Lewis index (T-L). This index (Tucker & Lewis, 1973) represents the ratio of the amount of variance associated with the model to the total variance; it may be interpreted as indicating how well a factor model with a given number of common factors represents the covariances among the parcels for a population of examinees. A low coefficient indicates that the relations among the parcels are more complex than can be represented by that number of common factors. The T-L index can be interpreted as a reliability coefficient, with increases across models indicating the percentage of variance gained by moving to a more complex model.

The chi-squared/df ratio. This index is based on the overall chi-square goodness of fit test associated with each factor model. Ratios up to 5.0 indicate a reasonable fit (Marsh & Hocevar, 1985). It should be noted that the ratio depends on sample size; however, this is not a problem where comparisons are between factor models, as in the present study.

Results

Comparison of Domestic and Overseas Examinees

The first set of analyses were those involving the comparison between examinees taking the TOEFL in domestic and in overseas test centers. The data used in these analyses were taken from the November 1984 TOEFL. The data for the domestic examinees were the same as those used in the study by Hale et al. (1988) except (a) the sample of Chinese speakers excluded certain subgroups, as indicated in the Method section, to make this sample comparable to that used by Swinton and Powers (1980), and (b) reconstitution of the data set resulted in small changes in numbers of subjects in the other language groups.

Mean performance. For each student, the mean proportion of correct items was computed for each subscore and for the total test (i.e., the number of items answered correctly divided by the number of items available). Proportion scores rather than mean number correct are reported, in order to facilitate comparison of item types with respect to difficulty. Means and standard deviations of the proportion-correct scores are presented in Table 1 for each language group, separately for domestic and overseas examinees.

For both domestic and overseas examinees, the rank ordering of groups from highest to lowest in total TOEFL scores was Spanish, Chinese, Farsi, Japanese, and Arabic. The ordering of these language groups was the same as that observed by Swinton and Powers (1980), except that in the latter study Farsi speakers scored lowest of the five groups. Apparently, the population of Farsi speakers seeking to study in the United States changed substantially as a result of the political situation in the early 1980s. Most important for the present analysis, however, is the fact that the rank-ordering of language groups proved to be the same for domestic and overseas groups. In this respect, then, there is a basic similarity in the populations of examinees taking the TOEFL in the United States or Canada and in other countries.

Correlational data. Table 2 presents, for each language group in domestic and overseas test centers, intercorrelations among the subscores of the TOEFL, the reliabilities of the scores, and the correlations corrected for attenuation (i.e., corrected for unreliability). Focusing on the corrected correlations, there was a relatively strong relationship between Sections 2 and 3 of the TOEFL--that is, (a) Structure and Written Expression and (b) Vocabulary and Reading Comprehension. The correlation between Section 1 (Listening Comprehension) and each of these other sections was less strong. This same general pattern appeared to be characteristic of every language group for both domestic and overseas examinees. Thus, the correlational data suggest that Sections 2 and 3 of the test measure processes that are highly related, while the Listening Comprehension section is somewhat distinct from the other two. This pattern is consistent with the factor-analytic results to be discussed below.

Table 1

Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the November 1984 TOEFL for Each Language Group, Tested in Domestic and Overseas Centers[a]

Language Group	N	List. Comp.	Struc. & Writ. Exp.	Vocab. & Read. Comp.	TOTAL
Domestic centers					
Arabic	1819	.60 (.19)	.62 (.19)	.49 (.18)	.56 (.17)
Chinese	2923	.66 (.16)	.69 (.15)	.63 (.16)	.65 (.14)
Farsi	477	.67 (.19)	.69 (.19)	.56 (.19)	.63 (.18)
Japanese	918	.58 (.19)	.66 (.18)	.53 (.19)	.58 (.17)
Spanish	1289	.70 (.20)	.73 (.18)	.68 (.16)	.70 (.16)
Overseas centers					
Arabic	1689	.55 (.21)	.65 (.18)	.53 (.18)	.57 (.17)
Chinese	2437	.63 (.20)	.73 (.17)	.68 (.17)	.68 (.17)
Farsi	105	.68 (.23)	.75 (.17)	.60 (.19)	.67 (.19)
Japanese	2037	.58 (.20)	.73 (.17)	.60 (.18)	.63 (.17)
Spanish	1280	.71 (.20)	.80 (.15)	.75 (.15)	.75 (.15)

[a] Scores are proportions correct (i.e., number correct divided by possible items) for each section: (1) Listening Comprehension, (2) Structure and Written Expression, and (3) Vocabulary and Reading Comprehension, and for the total test.

Table 2a

Domestic Test Centers:
Correlations Among TOEFL Subscores and Subscore Reliabilities for
Each Language Group (November 1984 TOEFL)[a]

		List.	Struc. &	Vocab. &
		Comp.	Writ. Exp.	Read. Comp.
Language Group		(LC)	(SWE)	(VRC)
Arabic	LC	[.90]	.82	.82
	SWE	.73	[.87]	.93
	VRC	.74	.83	[.90]
Chinese	LC	[.87]	.78	.83
	SWE	.66	[.82]	.91
	VRC	.73	.77	[.88]
Farsi	LC	[.91]	.82	.85
	SWE	.73	[.89]	.93
	VRC	.77	.83	[.91]
Japanese	LC	[.90]	.81	.85
	SWE	.71	[.86]	.93
	VRC	.77	.83	[.91]
Spanish	LC	[.92]	.85	.83
	SWE	.77	[.88]	.91
	VRC	.75	.80	[.89]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Table 2b

Overseas Test Centers:
Correlations Among TOEFL Subscores and Subscore Reliabilities for Each Language Group (November 1984 TOEFL)[a]

Language Group		List. Comp. (LC)	Struc. & Writ. Exp. (SWE)	Vocab. & Read. Comp. (VRC)
Arabic	LC	[.91]	.78	.83
	SWE	.69	[.87]	.92
	VRC	.75	.81	[.90]
Chinese	LC	[.91]	.82	.86
	SWE	.73	[.86]	.93
	VRC	.78	.83	[.91]
Farsi	LC	[.94]	.89	.84
	SWE	.81	[.88]	.93
	VRC	.78	.84	[.92]
Japanese	LC	[.91]	.79	.83
	SWE	.71	[.87]	.93
	VRC	.75	.83	[.91]
Spanish	LC	[.93]	.82	.78
	SWE	.73	[.86]	.89
	VRC	.71	.78	[.89]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Factor analysis. A set of confirmatory factor analyses was conducted for each language group, separately for domestic and overseas examinees. Variance-covariance matrices provided the input to the analyses. Analyses were performed separately for the domestic and overseas examinees, to determine whether a different factor structure is associated with these two different populations.

Table 3 presents the goodness of fit indicators for the one-, two-, and three-factor solutions separately for the domestic examinees (Table 3a) and overseas examinees (Table 3b). As expected, the GFI and Tucker-Lewis indices increased across models, and the RMSR and chi-squared/df indices decreased across models. Most important is the degree of change in values across models. It is clear that the indices were not at maximal values for the one-factor model (a rule of thumb for the GFI and Tucker-Lewis indices, for example, is that their values should be in the .90s before one can conclude that the associated model adequately fits the data), and that they changed markedly when going from a one- to a two-factor model. Thus, a two-factor solution apparently provides a markedly better fit to the data than does a one-factor solution. The indices also differed slightly between the two- and three-factor solutions, but the differences were so small, and the indices so near their limits with the two-factor solution, that two factors appear sufficient to account for performance.

Although there were slight variations across language groups, the basic pattern of results just described was observed for each language group, indicating a strong consistency in this pattern. Especially important for the present purposes is the fact that this same basic pattern of results was observed for both the domestic and overseas examinees. Thus, contrary to the first hypothesis presented in the Introduction, there does not appear to be a fundamental difference in number of factors underlying TOEFL performance for examinees tested in domestic test centers and for those tested overseas. (The same pattern of results was also observed when the domestic and overseas populations were combined, as discussed below.)

In sum, two factors appear sufficient to account for performance. The Listening Comprehension factor is somewhat distinct from factors related to the other sections of the test. And this seems to be true across language groups for both the domestic and overseas populations of examinees.

Further information about the factor structure of the tests is available in the factor loadings and intercorrelations among factors. These data were examined for the two-factor model, since this particular model appeared to provide the best fit, according to the analyses discussed above. The factor loadings for domestic and overseas examinees on the November 1984 TOEFL are presented in Appendix A, Tables A1 and A2. (Loadings are constrained to be zero or nonzero, depending on the portion of the test by which a factor is defined.) The fact that all nonzero loadings were significant and substantial in magnitude shows that there was highly reliable assessment of individua' differences on both factors. The magnitudes of loadings were comparable across language groups and across domestic and overseas populations, indicating little difference among populations in the strength of the listening and nonlistening factors.

Table 3a

Domestic Test Centers:
Indices from Confirmatory Factor Analyses of the TOEFL
for One-, Two-, and Three-Factor Solutions (November 1984 TOEFL)

	Language Group				
Index/model	Arabic	Chinese	Farsi	Japanese	Spanish
<u>Goodness of fit (GFI)</u>					
One factor	.89	.91	.86	.87	.86
Two factor	.97	.97	.94	.96	.95
Three factor	.98	.99	.95	.97	.98
<u>Root mean square</u>					
One factor	.35	.25	.33	.35	.29
Two factor	.11	.08	.15	.13	.11
Three factor	.10	.07	.15	.11	.09
<u>Tucker-Lewis</u>					
One factor	.90	.90	.89	.89	.89
Two factor	.97	.97	.96	.97	.96
Three factor	.98	.99	.97	.98	.99
<u>Chi-squared/df ratio</u>					
One factor	17.32	21.60	5.81	10.18	14.44
Two factor	5.34	6.53	2.56	3.58	5.20
Three factor	3.67	3.36	2.12	2.60	2.56

Table 3b

Overseas Test Centers:
Indices from Confirmatory Factor Analyses of the TOEFL
for One-, Two-, and Three-Factor Solutions (November 1984 TOEFL)

	Language Group				
Index/model	Arabic	Chinese	Farsi	Japanese	Spanish
Goodness of fit (GFI)					
One factor	.85	.89	.75	.85	.81
Two factor	.95	.96	.85	.95	.93
Three factor	.97	.98	.86	.96	.97
Root mean square					
One factor	.40	.30	.36	.39	.36
Two factor	.15	.11	.24	.12	.13
Three factor	.14	.10	.22	.11	.12
Tucker-Lewis					
One factor	.87	.91	.83	.87	.83
Two factor	.96	.97	.91	.96	.95
Three factor	.97	.98	.92	.97	.97
Chi-squared/df ratio					
One factor	20.23	21.89	2.86	25.21	20.53
Two factor	6.73	7.42	1.96	8.56	7.10
Three factor	5.03	4.75	1.90	6.31	4.02

Correlations between the two factors, presented in Appendix B, ranged from .82 to .89 across language groups and across domestic and overseas populations. (Note that correlations between factors are, in effect, correlations corrected for attenuation, or unreliability.) The fact that the correlations were below .90, and the fact that the goodness of fit indicators pointed to a two-factor solution, indicate that the listening and nonlistening parts of the test do appear to measure somewhat different aspects of proficiency. At the same time, the fact that the correlations were relatively high suggests that there is a reasonably strong relation between factors, due to the influence of overall English proficiency on performance in all sections of the test. Comparisons of correlations across populations suggested no consistent population differences in the relationship between factors. Thus, the correlational data, together with the factor loadings and goodness of fit indices, indicate a similarity across language groups and across domestic and overseas populations in the factor structure of the test.

Comparison of November 1976 and November 1984 Tests

In a second set of analyses, data from the November 1976 and November 1984 TOEFL administrations were examined. The combined population of domestic and overseas examinees was used in each case.

Mean performance. Table 4 presents the mean proportion correct (and SD) for each language group, for both the November 1976 and November 1984 forms of the TOEFL. In both test administrations, Spanish speakers scored highest. The Arabic and Japanese speakers were the lowest scoring groups on the November 1984 test, whereas the Farsi speakers scored lower than these two groups on the November 1976 test. The nature of the Farsi-speaking population taking the TOEFL thus changed during this time period, as mentioned above.

Table 4

Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the TOEFL for Each Language Group, for November 1976 and November 1984 Tests (Domestic and Overseas Centers Combined)[a]

Language Group	N	List. Comp.	Struc. & Writ. Exp.	Vocab. & Read. Comp.	TOTAL
November 1976 test					
Arabic	686	.65 (.19)	.59 (.18)	.58 (.17)	.61 (.17)
Chinese	998	.63 (.18)	.63 (.16)	.62 (.16)	.63 (.15)
Farsi	987	.55 (.20)	.50 (.17)	.49 (.16)	.51 (.16)
Japanese	997	.61 (.17)	.58 (.16)	.61 (.17)	.60 (.15)
Spanish	991	.71 (.20)	.64 (.19)	.74 (.15)	.70 (.16)
November 1984 test					
Arabic	3508	.58 (.20)	.64 (.19)	.51 (.18)	.56 (.17)
Chinese	5360	.64 (.18)	.71 (.16)	.65 (.17)	.66 (.15)
Farsi	582	.67 (.20)	.70 (.19)	.57 (.19)	.64 (.18)
Japanese	2955	.58 (.20)	.71 (.18)	.58 (.18)	.61 (.17)
Spanish	2569	.70 (.20)	.77 (.17)	.71 (.16)	.72 (.16)

[a] Scores are proportions correct (i.e., number correct divided by possible items) for each section: (1) Listening Comprehension, (2) Structure and Written Expression, and (3) Vocabulary and Reading Comprehension, and for the total test.

Correlational data. Table 5 presents, for each language group given the November 1976 test (Table 5a) and November 1984 test (Table 5b), intercorrelations among TOEFL subscores, reliabilities, and correlations corrected for attenuation. From the corrected correlations it is again apparent that there was a relatively strong relationship between TOEFL Sections 2 and 3 (Structure and Written Expression, Vocabulary and Reading Comprehension), whereas there was a less strong relation between Listening Comprehension and each of these two sections. Thus, Sections 2 and 3 appear to measure processes that are more highly related to each other than they are to those measured in the Listening Comprehension section.

Factor analysis. Confirmatory factor analyses were conducted separately for the November 1976 and November 1984 test administrations. Variance-covariance matrices provided the data for analysis.

Table 6 presents the goodness of fit indicators for the one-, two-, and three-factor solutions separately for the November 1976 test (Table 6a) and the November 1984 test (Table 6b). Once again, the indices changed markedly when going from a one-factor to a two-factor solution, but the changes observed when going from a two- to a three-factor solution were so small that two factors appear sufficient to account for performance. This pattern of results was observed across language groups. And most important for the present purposes, this pattern was observed for both the November 1976 and November 1984 test administrations. Thus, in both cases a two-factor solution seems to characterize performance on the TOEFL, where one factor is defined by the Listening Comprehension section, and the other factor is defined by the other sections of the test.

Factor loadings, which appear in Tables A3 and A4 of Appendix A, showed a consistent pattern across language groups and across the 1976 and 1984 tests. Correlations between factors were in the .80s for all groups, and differences among groups were too small to be of interest. In general, all aspects of the data pointed to a basic similarity among populations with regard to the test's factor structure.

Table 5a

November 1976 TOEFL:
Correlations Among Subscores and Subscore Reliabilities for Each Language Group (Combined Domestic and Overseas Populations)[a]

Language Group		List. Comp. (LC)	Struc. & Writ. Exp. (SWC)	Vocab. & Read. Comp. (VRC)
Arabic	LC	[.91]	.84	.82
	SWE	.74	[.85]	.90
	VRC	.74	.79	[.90]
Chinese	LC	[.89]	.84	.82
	SWE	.72	[.83]	.95
	VRC	.73	.81	[.89]
Farsi	LC	[.90]	.85	.80
	SWE	.73	[.82]	.92
	VRC	.71	.77	[.86]
Japanese	LC	[.88]	.80	.80
	SWE	.68	[.82]	.94
	VRC	.71	.81	[.89]
Spanish	LC	[.92]	.90	.84
	SWE	.81	[.87]	.91
	VRC	.76	.80	[.89]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Table 5b

November 1984 TOEFL:
Correlations Among Subscores and Subscore Reliabilities for Each Language Group (Combined Domestic and Overseas Populations)[a]

Language Group		List. Comp. (LC)	Struc. & Writ. Exp. (SWC)	Vocab. & Read. Comp. (VRC)
Arabic	LC	[.91]	.78	.79
	SWE	.69	[.87]	.93
	VRC	.72	.82	[.90]
Chinese	LC	[.89]	.78	.82
	SWE	.68	[.85]	.92
	VRC	.73	.80	[.90]
Farsi	LC	[.92]	.82	.84
	SWE	.74	[.89]	.93
	VRC	.77	.83	[.91]
Japanese	LC	[.91]	.78	.82
	SWE	.69	[.87]	.93
	VRC	.74	.83	[.91]
Spanish	LC	[.92]	.82	.79
	SWE	.74	[.88]	.90
	VRC	.72	.80	[.90]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Table 6a

November 1976 TOEFL:
Indices from Confirmatory Factor Analyses
for One-, Two-, and Three-Factor Solutions
(Combined Domestic and Overseas Populations)

	Language Group				
Index/model	Arabic	Chinese	Farsi	Japanese	Spanish
Goodness of fit (GFI)					
One factor	.85	.89	.87	.88	.88
Two factor	.93	.95	.93	.96	.96
Three factor	.95	.96	.97	.96	.97
Root mean square					
One factor	.32	.28	.24	.33	.30
Two factor	.15	.12	.13	.11	.13
Three factor	.14	.11	.11	.10	.11
Tucker-Lewis					
One factor	.98	.98	.98	.97	.98
Two factor	.99	.99	.99	.99	.99
Three factor	.99	.99	1.00	.99	1.00
Chi-squared/df ratio					
One factor	8.81	9.45	10.60	10.26	9.40
Two factor	4.36	4.17	5.85	3.86	3.53
Three factor	3.15	3.78	3.16	3.15	2.77

Table 6b

November 1984 TOEFL:
Indices from Confirmatory Factor Analyses
for One-, Two-, and Three-Factor Solutions
(Combined Domestic and Overseas Populations)

	Language Group				
Index/model	Arabic	Chinese	Farsi	Japanese	Spanish
Coodness of fit (GFI)					
One factor	.85	.88	.85	.85	.82
Two factors	.96	.97	.94	.95	.94
Three factors	.98	.98	.95	.97	.97
Root mean square					
One factor	.46	.32	.34	.43	.30
Two factor	.12	.10	.15	.13	.13
Three factor	.11	.09	.14	.11	.11
Tucker-Lewis					
One factor	.98	.98	.98	.98	.97
Two factor	.99	.99	.99	.99	.99
Three factor	1.00	1.00	1.00	1.00	1.00
Chi-squared/df ratio					
One factor	42.68	49.32	7.30	37.85	37.91
Two factor	10.69	13.36	2.84	11.14	11.56
Three factor	7.51	7.66	2.26	7.90	5.89

Comparison of Low- and High-Proficiency Examinees

Definition of the sample. Analyses of the November 1984 test data were performed for four language groups, Arabic, Chinese, Japanese, and Spanish. The Farsi group was excluded as it contained too few examinees for this analysis.

One parcel from each subsection of the test was used to comprise a measure of proficiency, consisting of 51 total items. Low- and high-proficiency groups were defined on the basis of this 51-item test. Factor analyses for the low- and high-proficiency groups were then performed on the 95 items in the remaining parcels, which will be called the "abbreviated TOEFL." While use of an abbreviated test admittedly leads to some reduction in representativeness, this disadvantage was believed to be outweighed by the advantage of defining proficiency level on the basis of a test that was independent of that used in the factor analyses.

Low- and high-proficiency groups were defined for each language group separately, in the following manner. For a given language group, mean performance on the 51-item test was determined. Then the score midway between the mean and the score of the lowest scoring examinee in that language group was calculated, as was the score midway between the mean and the score of the highest scoring examinee in that group. These two scores served as midpoints of distributions to be constructed around them for that language group. Examinees were randomly discarded to yield approximate normal distributions around these two scores, with a range of approximately two standard deviations on either side of each score. The shape of the curves did not deviate significantly from normality according to the Kolmogorov-Smirnov index (McNemar, 1962).

It should be noted that application of these procedures produced fewer examinees in the low- than the high-proficiency group for each language. This was due to skewness in the distribution for the entire population of TOEFL takers, such that larger numbers of examinees were available to produce a normal distribution in the high- than the low-proficiency group. Nevertheless, there were sufficient numbers of examinees in both low- and high-proficiency groups per language to perform factor analyses separately for each group.

Mean performance. Table 7 shows the mean proportion correct for the low- and high-proficiency examinees for each language. Given that the two proficiency groups were defined separately for each language, the ordering of languages within both of these groups paralleled that for the overall population of examinees, as would be expected.

Table 7

Means (and Standard Deviations in Parentheses) of Proportion-Correct Scores on the Abbreviated 1984 TOEFL for Each Language Group, for Low- and High-Proficiency Examinees[a]
(Combined Domestic and Overseas Populations)

Language Group	N	List. Comp.	Struc. & Writ. Exp.	Vocab. & Read. Comp.	TOTAL
Low proficiency					
Arabic	332	.38 (.13)	.44 (.15)	.32 (.11)	.37 (.10)
Chinese	434	.42 (.14)	.51 (.15)	.45 (.14)	.46 (.12)
Japanese	308	.38 (.14)	.49 (.16)	.37 (.12)	.40 (.11)
Spanish	230	.44 (.16)	.50 (.17)	.47 (.14)	.47 (.12)
High proficiency					
Arabic	923	.77 (.14)	.82 (.12)	.69 (.14)	.75 (.11)
Chinese	1659	.80 (.11)	.84 (.10)	.81 (.10)	.81 (.08)
Japanese	861	.78 (.14)	.87 (.09)	.77 (.11)	.80 (.09)
Spanish	834	.85 (.11)	.89 (.13)	.82 (.10)	.85 (.08)

[a] Proficiency groups are defined by performance on 51 test items; scores reported in the table are based on the remaining 95 test items. Scores are proportions correct (i.e., number correct divided by possible items) for each section: (1) Listening Comprehension, (2) Structure and Written Expression, and (3) Vocabulary and Reading Comprehension, and for the total test.

Correlational data. Table 8 presents the correlations among subscores, reliabilities, and corrected correlations; data for low-proficiency examinees are in Table 8a, and data for high-proficiency examinees are in Table 8b. As with the total populations for each of these four languages, the correlations between TOEFL Sections 2 and 3 were consistently higher than were the correlations between Listening Comprehension (Section 1) and each of the two other sections. This was true for both low- and high-proficiency examinees for each language, suggesting a basic similarity between proficiency groups in this respect.

It will also be observed that the Listening Comprehension section tended to correlate more highly with Section 3 (Vocabulary and Reading Comprehension) than with Section 2 (Structure and Written Expression). Nevertheless, the factor analyses presented below provide little indication that Sections 2 and 3 are associated with separate factors.

Factor analysis. Confirmatory factor analyses were conducted separately for low- and high-proficiency examinees in each language group. In comparing populations, it is always preferable to analyze variance-covariance matrices, as was done with the factor analyses reported above. Unfortunately, in this case the use of variance-covariance matrices led to out-of-bounds estimates and/or lack of convergence. Therefore, correlation matrices were used as the data for analyses here, and in these analyses there was convergence with acceptable estimates.

In examining the goodness of fit indicators for low- and high-proficiency examinees, it must be noted that the chi-squared/df ratio is the least meaningful. This index is sample-size dependent, and there was a substantial difference in numbers of examinees in the low- and-high proficiency groups. Furthermore, a chi-square statistic based on the correlation matrix (as used in this analysis) rather than the variance-covariance matrix is not entirely appropriate for interpretation. It will also be noted that the RMSR indices are considerably lower than were those observed in the previous analyses; this was due to the use of correlation matrices rather than variance-covariance matrices as the data for analysis.

Table 9 presents the results of these analyses, Table 9a for the low-proficiency examinees, and Table 9b for the high-proficiency examinees. For both low- and high-proficiency examinees, the goodness of fit indicators changed noticeably when going from a one- to a two-factor solution but very little when going from a two- to a three-factor solution. Thus, for both proficiency groups, no more than two factors appear to be required to account for performance. Indeed, the figures indicated quite good fit even with a single-factor solution. Still, the fit of the two-factor solution appeared to be sufficiently better than that of the one-factor solution to warrant concluding that, for both low- and high-proficiency examinees, the Listening Comprehension factor is somewhat distinct from the factors defined by the other test sections.

Table 8a

Low-Proficiency Examinees:
Correlations Among Subscores and Subscore Reliabilities for
Each Language Group on the Abbreviated 1984 TOEFL[a]
(Combined Domestic and Overseas Populations)

Language Group		List. Comp. (LC)	Struc. & Writ. Exp. (SWE)	Vocab. & Read. Comp. (VRC)
Arabic	LC	[.65]	.51	.53
	SWE	.35	[.72]	.81
	VRC	.35	.56	[.65]
Chinese	LC	[.71]	.57	.65
	SWE	.41	[.73]	.78
	VRC	.47	.58	[.76]
Japanese	LC	[.69]	.50	.59
	SWE	.36	[.73]	.81
	VRC	.40	.57	[.68]
Spanish	LC	[.76]	.53	.58
	SWE	.41	[.76]	.77
	VRC	.44	.59	[.77]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Table 8b

High-Proficiency Examinees:
Correlations Among Subscores and Subscore Reliabilities for
Each Language Group on the Abbreviated 1984 TOEFL[a]
(Combined Domestic and Overseas Populations)

Language Group		List. Comp. (LC)	Struc. & Writ. Exp. (SWE)	Vocab. & Read. Comp. (VRC)
Arabic	LC	[.78]	.50	.65
	SWE	.38	[.75]	.81
	VRC	.51	.63	[.80]
Chinese	LC	[.71]	.40	.59
	SWE	.28	[.70]	.74
	VRC	.42	.52	[.71]
Japanese	LC	[.78]	.40	.67
	SWE	.29	[.66]	.77
	VRC	.51	.54	[.73]
Spanish	LC	[.74]	.51	.50
	SWE	.36	[.66]	.74
	VRC	.37	.51	[.72]

[a] Coefficient alpha reliabilities are in brackets along the diagonal for each language group. Raw correlations appear below the diagonal, and correlations corrected for attenuation appear above the diagonal.

Table 9a

Low-Proficiency Examinees:
Indices from Confirmatory Factor Analyses of the Abbreviated 1984 TOEFL for One-, Two-, and Three-Factor Solutions
(Combined Domestic and Overseas Populations)

	Language Group			
Index/model	Arabic	Chinese	Japanese	Spanish
Goodness of fit (GFI)				
One factor	.94	.94	.91	.91
Two factor	.97	.97	.97	.96
Three factor	.97	.98	.97	.97
Root mean square				
One factor	.06	.05	.07	.07
Two factor	.04	.04	.04	.04
Three factor	.04	.03	.04	.04
Tucker-Lewis				
One factor	.93	.93	.84	.91
Two factor	1.00	1.00	1.00	1.00
Three factor	1.00	1.00	1.00	1.00
Chi-squared/df ratio				
One factor	1.28	1.58	1.74	1.39
Two factor	0.58	0.71	0.59	0.57
Three factor	0.56	0.60	0.56	0.46

Table 9b

High-Proficiency Examinees:
Indices from Confirmatory Factor Analyses of the Abbreviated
1984 TOEFL for One-, Two-, and Three-Factor Solutions
(Combined Domestic and Overseas Populations)

	Language Group			
Index/model	Arabic	Chinese	Japanese	Spanish
Goodness of fit (GFI)				
One factor	.91	.92	.91	.91
Two factor	.97	.97	.97	.96
Three factor	.98	.98	.98	.97
Root mean square				
One factor	.06	.07	.07	.07
Two factor	.04	.04	.04	.05
Three factor	.03	.03	.03	.04
Tucker-Lewis				
One factor	.83	.78	.79	.74
Two factor	.98	.94	.97	.95
Three factor	1.00	.98	1.00	.97
Chi-squared/df ratio				
One factor	5.26	7.56	4.64	4.78
Two factor	1.51	2.73	1.48	1.79
Three factor	0.89	1.65	0.87	1.38

Factor loadings are presented in Tables A5 and A6 of Appendix A; correlations between factors are shown in Appendix B. (Factor loadings are less than 1 in this case, due to the use of the correlation matrix rather than the covariance matrix as input to the analyses; correlations are lower than those for other populations due to the restriction in range for low- and high-proficiency examinees.) The factor loadings are comparable across language groups and across low- and high-proficiency groups, and the correlations show no consistent differences across these groups. Thus, these data support the conclusion that there is little difference in factor structure of the test for these various populations.

Discussion

A recent study by Hale et al. (1988) suggested that TOEFL performance can be characterized by two factors for each of several language groups, whereas a previous study by Swinton and Powers (1980) had suggested a three-factor structure. The present study extended the research of Hale et al. in order to investigate possible bases for the inconsistency in conclusions reached in these two studies.

Comparison of Domestic and Overseas Examinees

One hypothesis was that the inconsistency resulted from use of different examinee populations: Hale et al. used only domestic examinees, whereas Swinton and Powers used the combined overseas and domestic population. To address this hypothesis, confirmatory factor analyses similar to those used by Hale et al. were performed on TOEFL data for both the overseas and domestic populations, as well as for the combined population. The hypothesis was not borne out, as analyses for the overseas and combined populations, as well as for the domestic examinees, produced comparable results.

In all cases, TOEFL performance appeared to be characterized by two factors, one defined by the Listening Comprehension section, and the other defined by the nonlistening sections of the test, a finding that was consistent across language groups. The relatively high correlation between factors for each population suggests that the empirical distinction between these factors is not strong. Nevertheless, the degree of fit was sufficiently greater for the two- than the one-factor solution to warrant a two-factor interpretation for each of the populations under study.

The similarity in results for domestic and overseas examinees has important practical implications. As noted above, some TOEFL research is conducted with domestic examinees only, under the assumption that those examinees are reasonably representative of the total population of TOEFL takers. To date, support for this assumption has come from statistical analyses showing similarities in domestic and overseas populations with regard to certain basic item statistics. The present study provided the only comparison thus far of the factor structure of the TOEFL for these two populations. The fact that the results appeared to be roughly the same in both cases suggests that the factor structure of the test is similar for both populations. In this respect, the domestic and overseas populations appear to be comparable.

Comparison of November 1976 and November 1984 Tests

Another difference between the Swinton and Powers study and that by Hale et al. lay in the methods used for data analysis. Hale et al. used confirmatory factor analysis following the procedures of LISREL VI, with factors defined by the sections of the test, whereas Swinton and Powers used other methods (see footnote 1 in the Introduction). Furthermore, Hale et al.

used item parcels as the units of analysis, whereas Swinton and Powers used individual items, and it has been argued that use of individual items can sometimes produce a factor associated with item difficulty (Hulin, Drasgow, & Parsons, 1983).

In the present study, data from the Swinton and Powers study (the November 1976 test) were reanalyzed, using parcels as the units of analyses and LISREL confirmatory factor-analytic methods. If the results were to suggest a factor structure different from that implied by Swinton and Powers' data, it could be inferred that differences in factor-analytic methodology played a role. Indeed, confirmatory factor analyses of the November 1976 data, conducted separately for each of the five language groups, suggested an interpretation different from that offered by Swinton and Powers and similar to that observed for the November 1984 data: performance on the 1976 as well as the 1984 test can be accounted for by two factors at most. Apparently, differences in factor-analytic methodology did contribute to the different conclusions reached in the present study and that of Swinton and Powers.

It was not within the scope of this study to determine exactly which aspects of the methods contributed to the inconsistency in interpretations drawn in these two studies. The principal conclusion to be drawn here is that, when confirmatory factor analysis is used, and factors are defined by the sections of the TOEFL, the results support the interpretation that TOEFL performance can best be characterized by two factors (albeit relatively highly related factors). One of these is defined by the Listening Comprehension section, and the other is defined by the remaining, nonlistening sections of the test.

An important implication of these results is that the factor structure of the TOEFL appears not to have changed between 1976 and 1984. If application of the same factor-analytic methods with the two test forms had produced different results, it might have been concluded that there had been a change over time in the basic structure of the TOEFL. That this was not the case, however, suggests that the test development process has not undergone any major evolution over this time period. Of course, a more definitive statement in this regard would require research comparing 1976 and 1984 test forms with samples from the same cohort--that is, samples drawn within the same time frame. Tentatively, however, it appears unlikely that the different results obtained by Swinton and Powers and Hale et al. can be attributed to changes in the factor structure of the test over time.

Comparison of Low- and High-Proficiency Examinees

An additional objective of the study was to assess the role of proficiency in TOEFL's factor structure. The results indicated that a two-factor solution best accounted for performance, and that this appeared to be the case for both low- and high-proficiency examinees. Thus, the results supported neither the hypothesis of increasing factor differentiation, nor that of decreasing differentiation, as a function of increasing examinee proficiency. Rather, the data suggested, as with the total population, that TOEFL performance of examinees at either extreme of proficiency is characterized by two factors, one defined by the Listening Comprehension section, and the other, by the nonlistening sections.

This conclusion contrasts with that of Oltman et al. (1988), who, using multidimensional scaling and cluster analysis, observed greater differentiation among dimensions for low- than for medium- and high-proficiency examinees. One possible explanation for the difference in results has to do with the methods used. Oltman et al.'s analysis was designed in a way that allowed separate dimensional structures to emerge for easy and for difficult items, and they found that it was the easy items for which a more differentiated structure was observed. In the present study, on the other hand, the analysis was deliberately constrained in such a way that item difficulty could not play a role. Another possible explanation is that variances of proficiency groups were made approximately equal in the present study but not in the study by Oltman et al., and group differences in variability can produce apparent group differences in degree of differentiation among dimensions. Other aspects of methodology may also have played a role, and the basis for the difference in results cannot be resolved without further study.

Examination of Other Factor-Analytic Studies Involving the TOEFL

Studies using the current three-part TOEFL. Although the previous research by Swinton and Powers and Hale et al. provided the primary background for the present study, some other factor-analytic studies also contribute information about the factor structure of the TOEFL. Dunbar (1982), in a confirmatory factor analysis of Swinton and Powers' data, compared models involving one factor and four factors (a general factor plus one factor for each TOEFL section). In the latter model, the general factor appeared to relate particularly to the nonlistening sections. After this factor was partialed out, a fairly large factor related to Listening Comprehension remained, suggesting that the Listening Comprehension section is more distinct from the other two sections than the latter are from each other.

Manning (1987) examined TOEFL performance along with teachers' and students' self-ratings of proficiency. In an exploratory factor analysis with oblique (Promax) rotation and extraction of four factors, two of the factors related to TOEFL performance, one involving the Listening Comprehension section, and the other involving the nonlistening sections.

Davidson (1988) factor analyzed the TOEFL among other measures, using principal components analyses and two- and three-factor solutions, with oblique (Promax) rotations. The data pointed to a separate factor associated with the Listening Comprehension section but no clear distinction between the two nonlistening sections of the TOEFL.

Data from operational administrations of the TOEFL (Educational Testing Service, 1987) also deserve consideration, even though these data were subjected to correlational but not factor analysis. Across 12 administrations, the average disattenuated correlation (i.e., correlation corrected for unreliability) was .90 between the two nonlistening sections

(Structure and Written Expression; Vocabulary and Reading Comprehension), but only .77 and .76 between Listening Comprehension and each of the latter two sections, respectively.

All of these data suggest that the skills measured by the Listening Comprehension section of the TOEFL are somewhat distinct from those measured by the nonlistening sections, whereas the latter two sections are more highly related to each other than either is to Listening Comprehension.

Studies using the five-part TOEFL. Other factor-analytic studies involving the TOEFL used the original five-part test, which was employed in 1975 and earlier; the five parts were essentially comparable to the five subsections of the current test, although there were a few differences.

Two studies (Stevenson, 1975; Oller & Hinofotis, 1980), using foreign students of various language backgrounds and employing the TOEFL along with other measures, observed that TOEFL Listening Comprehension was associated with one factor, and all nonlistening sections with a second factor (although the first study did not attempt a solution involving more than two factors). Three studies, one with Iranian students (Oller & Hinofotis, 1980), one with Nigerian high school students (Osanyinbi, 1975), and one with ESL students of various language backgrounds (Hosley & Meredith, 1979), obtained results suggesting that a single factor is associated with TOEFL performance. A study of Thai educators (Gue & Holdaway, 1973) found that three factors could be identified that were linked to different groupings of TOEFL sections.

The different conclusions suggested by the studies just mentioned are undoubtedly related to the use of varying types of samples and varying factor-analytic methodologies, as well as to the use of very small numbers of examinees (between 51 and 217), which can affect the stability of the results obtained. Whatever the basis for the differences, however, the fact that these studies used the old five-part TOEFL, and the fact that they were not generally concerned with the internal structure of the TOEFL, render these studies only tangentially relevant to the objectives of the present research.

Conclusions and Practical Implications

The results of this study suggested that a two-factor model best characterized TOEFL performance for each of the populations studied, with the two factors defined by the Listening Comprehension section and the nonlistening sections. Although the distinction between these factors was not strong, which reflects the pervasive influence of general proficiency in English, the Listening Comprehension section still appeared to measure an aspect of performance that was, to some extent, different from that tapped by the nonlistening sections. The latter sections, on the other hand, appeared to measure aspects of performance that were not clearly separate from each other.

An important question concerns the implications of these results for use of the three TOEFL section scores. In particular, it is of value to examine the degree to which the section subscores have diagnostic potential, in the sense that they provide information about separate and distinct areas of English proficiency. While there appears to be at least some merit in differentiating between scores on the Listening Comprehension and nonlistening parts of the test, the diagnostic value of the distinction between nonlistening subscores is less apparent.

In evaluating the three-section structure of the TOEFL, however, other considerations besides factor-analytic results must also play a role. One important consideration involves item content. Conceptually distinct constructs are examined in the two nonlistening sections, (a) Structure and Written Expression, whose items test knowledge of grammatical structure and usage in written English, and (b) Vocabulary and Reading Comprehension, whose items test lexical knowledge and ability to comprehend English text. Thus, there appears to be a relatively solid content-related basis for distinguishing between these two areas of proficiency, as well as between these areas and listening comprehension, in construction of the test.

Perhaps the strong relation between the two nonlistening sections is due as much to current instructional practices as to inherent similarities in these two areas of proficiency, since English grammatical structure is commonly taught hand-in-hand with vocabulary and reading comprehension. Well-designed experimental research on instructional effects could likely clarify this matter. It might well be possible to show that the different types of nonlistening content would respond differentially to different instructional emphasis. If that is the case, the separate TOEFL section scores would clearly have diagnostic value to individual students and instructors as well. Such a study would be useful from a practical standpoint and could also add important evidence concerning the construct validity of the test.

The recent multidimensional scaling and clustering analysis by Oltman et al. (1988) also bears on the structure of the TOEFL. Those authors identified separate dimensions corresponding to the three TOEFL sections, where the dimensions were defined largely by the easy items in each section, and the differentiation among dimensions was most pronounced for low-proficiency examinees. Oltman et al.'s results are consistent with the proposition that experimental studies of instructional effects would better reveal distinctive constructs that correspond more clearly to TOEFL content. Examinees at early levels of acquisition are more likely to have had somewhat different learning experiences that would produce distinguishable correlational patterns.

In sum, the value of the current TOEFL structure appears to depend on several considerations. On the basis of the present correlational/factor-analytic evidence, the diagnostic potential of the distinction among TOEFL sections--at least the distinction between the nonlistening sections--does not appear to be substantial. From a content standpoint, on the other hand, there is a clear difference in the aspects of proficiency tapped by the three sections. Thus, there is a conceptual basis for distinguishing among the areas of proficiency covered by these three sections, whether or not the section scores have clearly demonstrable diagnostic value on the basis of correlational evidence.

The present investigation, together with that of Oltman et al., suggest some provocative areas for further study. Additional research that allows item difficulty to play a role, using correlational as well as experimental evidence, could provide a more comprehensive test of the hypothesis that the structure of the TOEFL is more differentiated for easy than difficult items. Also, continued research comparing effects for low- and for high-proficiency examinees would be useful, to resolve the apparent discrepancy in results of these two studies and to provide a further test of the hypothesis that subtest differentiation varies with examinee proficiency. Further, it would be helpful to know how the specific types of skill and knowledge tested by the TOEFL are responsive to an effective instructional program at different levels of acquisition.

An especially profitable direction, from a diagnostic viewpoint, might be to establish a scale of proficiency levels within each content area. Students may have the same rank ordering in the various content areas, thus ensuring a high correlation between content areas, yet a given student may perform below an acceptable threshold level of proficiency in one content area and above the threshold level in another content area. Scores based on assessment of proficiency levels may well prove to be have more diagnostic utility than the typical normative scores.

References

Anastasi, A. (1970). On the formation of psychological traits. *American Psychologist*, *25*, 899-910.

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

Dunbar, S. B. (1982). *Construct validity and the internal structure of a foreign language test for several native language groups*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Educational Testing Service (1987). *TOEFL test and score manual*. Princeton, NJ: Author.

Gue, L. R., & Holdaway, D. A. (1973). English proficiency tests as predictors of success in graduate studies in education. *Language Learning*, *23*, 89-103.

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, ETS Research Rep. No. 88-2). Princeton, NJ: Educational Testing Service.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher*. Skokie, IL: National Textbook Company.

Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly*, *13*, 209-217.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL V--Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago: International Educational Services.

Joreskog, K. G., & Sorbom, D. (1983). *LISREL VI--Supplement to the LISREL V manual*. Uppsala, Sweden: University of Uppsala, Department of Statistics.

Joreskog, K., & Sorbom, D. (1985). *LISREL VI--An analysis of linear structural relationships by the method of maximum likelihood*. Mooresville, IN: Scientific Software.

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

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

McNemar, Q. (1962). Psychological Statistics (3rd ed.). New York: Wiley.

Oller, J. W., Jr., & Hinofotis, F. B. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. W. Oller, Jr., & K. Perkins (Eds.), Research in language testing (pp. 13-23). Rowley, MA: Newbury House.

Oltman, P. K., Stricker, L. J., & Barrows, R. S. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language. (TOEFL Research Rep. No. 27, ETS Research Rep. No. 88-26). Princeton, NJ: Educational Testing Service.

Osanyinbi, J. A. (1975). A concurrent validity study of the West African School Certificate and General Certificate of Education English Language Examination, using Educational Testing Service's Test of English as a Foreign Language as the criterion measure (Doctoral dissertation, University of Wisconsin, 1974). Dissertation Abstracts International, 35, 5130A-5131A. (University Microfilms No. 74-22, 134)

Selinker, L. (1969). Language transfer. General Linguistics, 9, 67-92.

Stevenson, D. K. (1975). A preliminary investigation of construct validity and the Test of English as a Foreign Language (Doctoral disseration, University of New Mexico, 1974). Dissertation Abstracts International, 36, 1352A. (University Microfilms No. 75-18, 664)

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, ETS Research Rep. No. 80-32). Princeton, NJ: Educational Testing Service.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Appendix A

Factor Loadings for Two-Factor Solution

Table A1

Factor Loadings for Two-Factor Solution

November 1984 TOEFL--Domestic Centers

	Arabic Group		Chinese Group		Farsi Group		Japanese Group		Spanish Group	
	Factor		Factor		Factor		Factor		Factor	
Parcel	1	2	1	2	1	2	1	2	1	2
LC1	3.199	0.0	2.803	0.0	3.323	0.0	3.309	0.0	3.065	0.0
LC2	3.228	0.0	2.614	0.0	3.091	0.0	3.165	0.0	3.360	0.0
LC3	2.616	0.0	2.281	0.0	2.769	0.0	2.784	0.0	2.997	0.0
S1	0.0	1.080	0.0	0.889	0.0	1.233	0.0	1.138	0.0	1.050
S2	0.0	1.115	0.0	0.906	0.0	1.027	0.0	1.150	0.0	1.090
WE1	0.0	1.354	0.0	1.015	0.0	1.362	0.0	1.179	0.0	1.254
WE2	0.0	1.442	0.0	1.038	0.0	1.463	0.0	1.416	0.0	1.396
WE3	0.0	1.517	0.0	1.187	0.0	1.514	0.0	1.293	0.0	1.366
V1	0.0	1.908	0.0	1.608	0.0	1.813	0.0	1.889	0.0	1.447
V2	0.0	1.824	0.0	1.399	0.0	1.864	0.0	1.729	0.0	1.439
V3	0.0	1.662	0.0	1.395	0.0	1.577	0.0	1.536	0.0	1.020
RC1	0.0	1.641	0.0	1.424	0.0	1.776	0.0	1.757	0.0	1.609
RC2	0.0	1.469	0.0	1.333	0.0	1.653	0.0	1.778	0.0	1.499
RC3	0.0	1.446	0.0	1.294	0.0	1.611	0.0	1.656	0.0	1.641

Table A2

Factor Loadings for Two-Factor Solution

November 1984 TOEFL--Overseas Centers

	Arabic Group Factor		Chinese Group Factor		Farsi Group Factor		Japanese Group Factor		Spanish Group Factor	
Parcel	1	2	1	2	1	2	1	2	1	2
LC1	3.517	0.0	3.268	0.0	4.090	0.0	3.377	0.0	3.117	0.0
LC2	3.491	0.0	3.229	0.0	3.848	0.0	3.175	0.0	3.569	0.0.
LC3	2.865	0.0	2.832	0.0	3.267	0.0	2.882	0.0	3.146	0.0
S1	0.0	1.100	0.0	1.023	0.0	0.987	0.0	0.997	0.0	0.885
S2	0.0	1.045	0.0	0.992	0.0	0.893	0.0	1.213	0.0	0.925
WE1	0.0	1.316	0.0	1.171	0.0	1.335	0.0	1.163	0.0	1.013
WE2	0.0	1.298	0.0	1.228	0.0	1.260	0.0	1.331	0.0	1.251
WE3	0.0	1.468	0.0	1.314	0.0	1.463	0.0	1.238	0.0	1.105
V1	0.0	1.815	0.0	1.709	0.0	1.780	0.0	1.614	0.0	1.284
V2	0.0	1.783	0.0	1.623	0.0	1.830	0.0	1.728	0.0	1.318
V3	0.0	1.552	0.0	1.561	0.0	1.694	0.0	1.410	0.0	0.966
RC1	0.0	1.601	0.0	1.679	0.0	1.941	0.0	1.697	0.0	1.468
RC2	0.0	1.457	0.0	1.476	0.0	1.356	0.0	1.710	0.0	1.441
RC3	0.0	1.394	0.0	1.501	0.0	1.813	0.0	1.606	0.0	1.546

Table A3

Factor Loadings for Two-Factor Solution

November 1984 TOEFL--Domestic and Overseas Centers

	Arabic Group Factor		Chinese Group Factor		Farsi Group Factor		Japanese Group Factor		Spanish Group Factor	
Parcel	1	2	1	2	1	2	1	2	1	2
LC1	3.379	0.0	3.028	0.0	3.459	0.0	3.358	0.0	3.096	0.0
LC2	3.392	0.0	2.916	0.0	3.249	0.0	3.171	0.0	3.462	0.0
LC3	2.772	0.0	2.554	0.0	2.872	0.0	2.852	0.0	3.075	0.0
S1	0.0	1.094	0.0	0.966	0.0	1.201	0.0	1.070	0.0	1.003
S2	0.0	1.089	0.0	0.957	0.0	1.012	0.0	1.216	0.0	1.036
WE1	0.0	1.340	0.0	1.093	0.0	1.372	0.0	1.178	0.0	1.176
WE2	0.0	1.381	0.0	1.137	0.0	1.438	0.0	1.391	0.0	1.367
WE3	0.0	1.502	0.0	1.265	0.0	1.512	0.0	1.289	0.0	1.273
V1	0.0	1.889	0.0	1.685	0.0	1.806	0.0	1.751	0.0	1.441
V2	0.0	1.822	0.0	1.540	0.0	1.879	0.0	1.769	0.0	1.401
V3	0.0	1.619	0.0	1.487	0.0	1.609	0.0	1.466	0.0	0.990
RC1	0.0	1.628	0.0	1.560	0.0	1.809	0.0	1.743	0.0	1.594
RC2	0.0	1.469	0.0	1.415	0.0	1.600	0.0	1.752	0.0	1.509
RC3	0.0	1.429	0.0	1.402	0.0	1.654	0.0	1.638	0.0	1.648

Table A4

Factor Loadings for Two-Factor Solution

November 1976 TOEFL--Combined Domestic and Overseas Centers

	Arabic Group		Chinese Group		Farsi Group		Japanese Group		Spanish Group	
	Factor		Factor		Factor		Factor		Factor	
Parcel	1	2	1	2	1	2	1	2	1	2
LC1	2.907	0.0	2.755	0.0	3.197	0.0	2.639	0.0	3.209	0.0
LC2	3.337	0.0	2.885	0.0	3.228	0.0	2.622	0.0	3.040	0.0
LC3	2.940	0.0	2.724	0.0	2.932	0.0	2.793	0.0	3.086	0.0
S1	0.0	1.007	0.0	1.148	0.0	1.203	0.0	0.993	0.0	0.931
S2	0.0	1.006	0.0	1.113	0.0	1.003	0.0	0.963	0.0	0.988
WE1	0.0	1.087	0.0	0.956	0.0	1.200	0.0	1.165	0.0	1.083
WE2	0.0	1.403	0.0	1.138	0.0	1.594	0.0	1.197	0.0	1.315
WE3	0.0	1.463	0.0	1.200	0.0	1.523	0.0	1.131	0.0	1.266
V1	0.0	1.415	0.0	1.374	0.0	0.867	0.0	1.087	0.0	1.048
V2	0.0	1.361	0.0	1.295	0.0	1.379	0.0	1.152	0.0	0.961
V3	0.0	1.199	0.0	1.160	0.0	1.149	0.0	1.322	0.0	1.010
RC1	0.0	1.867	0.0	1.639	0.0	1.533	0.0	1.833	0.0	1.641
RC2	0.0	1.918	0.0	1.884	0.0	1.733	0.0	2.111	0.0	1.916
RC3	0.0	1.633	0.0	1.476	0.0	1.502	0.0	1.704	0.0	1.616

Table A5

Factor Loadings for Two-Factor Solution

November 1984 Abbreviated TOEFL--Low-Proficiency Examinees

	Arabic Group Factor		Chinese Group Factor		Japanese Group Factor		Spanish Group Factor	
Parcel	1	2	1	2	1	2	1	2
LC1	0.840	0.0	0.946	0.0	0.899	0.0	0.888	0.0
LC2	0.582	0.0	0.556	0.0	0.681	0.0	0.706	0.0
S1	0.0	0.451	0.0	0.551	0.0	0.620	0.0	0.495
WE1	0.0	0.619	0.0	0.551	0.0	0.636	0.0	0.601
WE2	0.0	0.627	0.0	0.562	0.0	0.524	0.0	0.640
V1	0.0	0.611	0.0	0.690	0.0	0.596	0.0	0.661
V2	0.0	0.548	0.0	0.539	0.0	0.579	0.0	0.600
RC1	0.0	0.387	0.0	0.568	0.0	0.452	0.0	0.571
RC2	0.0	0.393	0.0	0.598	0.0	0.422	0.0	0.621

Table A6

Factor Loadings for Two-Factor Solution

November 1984 Abbreviated TOEFL--High-Proficiency Examinees

	Arabic Group Factor		Chinese Group Factor		Japanese Group Factor		Spanish Group Factor	
Parcel	1	2	1	2	1	2	1	2
LC1	0.885	0.0	0.832	0.0	0.829	0.0	0.771	0.0
LC2	0.745	0.0	0.681	0.0	0.793	0.0	0.792	0.0
S1	0.0	0.501	0.0	0.480	0.0	0.473	0.0	0.408
WE1	0.0	0.553	0.0	0.471	0.0	0.460	0.0	0.521
WE2	0.0	0.708	0.0	0.639	0.0	0.541	0.0	0.593
V1	0.0	0.663	0.0	0.573	0.0	0.515	0.0	0.578
V2	0.0	0.728	0.0	0.629	0.0	0.565	0.0	0.524
RC1	0.0	0.660	0.0	0.522	0.0	0.670	0.0	0.582
RC2	0.0	0.671	0.0	0.589	0.0	0.679	0.0	0.620

Appendix B

Correlations Between Factors in Two-Factor Confirmatory Factor Analyses

Table B

Correlations Between Factors in Two-Factor Solution
(Factors Defined as Listening Comprehension
and All Other Parts of the TOEFL)

Language Group	Nov. 1984 Domestic Examinees	Nov. 1984 Overseas Examinees	Nov. 1984 Combined Populations	Nov. 1976 Combined Populations
Arabic	.84	.83	.80	.85
Chinese	.83	.87	.83	.84
Farsi	.85	.88	.85	.89
Japanese	.84	.83	.81	.81
Spanish	.86	.82	.82	.84

Language Group	November 1984 Low-Proficiency Group	November 1984 High-Proficiency Group
Arabic	.54	.63
Chinese	.65	.57
Japanese	.55	.63
Spanish	.61	.55

57906-01193 • HP99M.5 • 275633 • Pnnted in U S A.