DOCUMENT RESUME

ED 390 894 TM 024 176

AUTHOR Henning, Grant
TITLE Test-Retest Analyses of the Test of English as a Foreign Language. TOEFL Research Reports Report 45.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-31
PUB DATE Jun 93
NOTE 33p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *English (Second Language); *Error of Measurement; *Estimation (Mathematics); Language Proficiency; *Language Tests; *Scores; Second Language Learning; Test Format; Test Items; Test Length; Test Reliability; Test Results
IDENTIFIERS *Test of English as a Foreign Language; Test Retest Reliability

ABSTRACT

This study provides information about the total and component scores of the Test of English as a Foreign Language (TOEFL). First, the study provides comparative global and component estimates of test-retest, alternate-form, and internal-consistency reliability, controlling for sources of measurement error inherent in the examinees and the testing administration context. The study also provides information about differential changes in subtest difficulty on repeated application over a small interval of time (8 days). This study considered the phenomenon of "item bounce" and reflected the comparative stabilities of difficulty estimates within item type over repeated test administrations. Although test-length-adjusted reliability estimates were found to be adequately high across reported component and total test scores, the study contained several inherent limitations, chief of which was the comparatively small sample of 329 subjects. Attrition and design features reduced the test-retest reliability estimates to separate repeating subgroups of 101 and 91 persons, and alternate-form reliability estimates were based on 52 and 25 persons. In addition, samples were not perfectly representative of the current TOEFL population. (Contains 10 tables and 14 references.) (SLD)

TOEFL

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 45
JUNE 1993

Test-Retest Analyses of the Test of English as a Foreign Language

Grant Henning

ETS
Educational Testing Servic

Test-Retest Analyses of the Test of English as a Foreign Language

Grant Henning

Educational Testing Service
Princeton, New Jersey

RR-93-31

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

Abstract

The present study provides two kinds of information that have not previously been available in a single research report on the Test of English as a Foreign Language (TOEFL) with regard to its total and component scores. First, the study provides comparative global and component estimates of test-retest, alternate-form, and internal-consistency reliability. This complies with the joint standards of the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education by controlling for sources of measurement error that may be inherent both among the examinees and within the testing administration context, and not merely within the examination itself. Secondly, the study provides information about differential change in subtest difficulty on repeated application over a small interval of time (viz., eight days). This second concern is related to the phenomenon of "item bounce" and reflects the comparative stabilities of difficulty estimates within item type over repeated test administrations. This comparative stability information may provide useful insights into the functioning of particular TOEFL® subtest item types and the suitability of those item types for anchoring in test equating.

Although test-length-adjusted reliability estimates were found to be adequately high across reported component and total test scores, with raw score test-retest coefficients ranging from .87 to .98 (with a mean of .93 over 22 total coefficients), raw score internal-consistency coefficients ranging from .79 to .98 (with a mean of .94 over 88 total coefficients), and raw score alternate-form coefficients ranging from .78 to .97 (with a mean of .90 over 22 total coefficients), the study contained several inherent limitations. Chief among these limitations was the comparatively small sample involved. Only 329 total subjects participated, and, due to attrition and design features, test-retest reliability estimates were based on separate repeating subgroups of only 101 and 91 persons. Alternate-form reliability estimates were based on separate repeating subgroups of only 52 and 25 persons. Although estimates were replicated across two TOEFL forms and at least two distinct samples, it was not possible within the existing project constraints to identify repeating samples that were perfectly representative of the current TOEFL examinee population in regard to language background and mean language proficiency.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than 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 the TOEFL test 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. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1992-93) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins (Chair)	Southern Illinois University at Carbondale
Linda Schinke-Llano	Millikin University
John Upshur	Concordia University

Table of Contents

List of Tables

Introduction

Background

The American Psychological Association standards for educational and psychological testing, also adopted as joint standards by the American Educational Research Association and the National Council on Measurement in Education, clearly require that test developers report alternate-form and/or test-retest reliability estimates in addition to internal-consistency estimates for all objective tests whenever inconsistencies in examinee performance or administrative environments may constitute relevant sources of measurement error (Novick et al., 1985; Stiggins & Bridgeford, 1983). In the words of Standard 2.6,

"Coefficients based on internal analysis should not be interpreted as substitutes for alternate-form reliability or estimates of stability over time unless other evidence supports that interpretation in a particular context." (Novick et al., 1985; p. 21).

Internal-consistency reliability is routinely reported for every TOEFL test form. In addition, Wilson (1987) has studied patterns of mean score change on TOEFL repetition after intervals permitting instructional intervention. Prior to the present study, however, there has been no ETS-sponsored research directly addressing test-retest reliability of the TOEFL test. In part, this is understandable because it is difficult to locate reasonably large and representative samples of examinees to participate in the study by repeating tests after a short, but appropriate, interval of time. Also, with most language tests there is the potential problem of practice effect when the tests are administered repeatedly. It is not known to what extent TOEFL subtests may be differentially susceptible to practice effect depending on the exact nature of the respective item types. Due to the potential problems of practice effect, maturation, learning, or forgetting on the part of the examinees, internal-consistency and alternate-form reliability estimates may also be more easily interpreted than test-retest coefficients.

The basic problem at issue is that the currently available internal-consistency reliability estimation fails to address certain examinee- or situation-related sources of measurement error. Unreliability due to examinee fatigue, illness, temporary anxiety, mood shift, or environmental distractions or other changes in the testing administrative conditions over time is not adequately reflected in estimates of internal-consistency reliability (Henning, 1987; Magnusson, 1966; Novick et al., 1985). A research question of interest arises from the consideration that, for some components or subtests of the TOEFL test, acceptable levels of internal-consistency reliability could be present while test-retest estimates might be less than adequate.

A related consideration is that the discrepancy between internal-consistency estimates and test-retest estimates is likely to vary depending on the nature of skills assessed or item formats employed, even when the number of items in the test or the amount of time spent on tasks are held constant. It would be of particular interest to identify testing components, tasks, or methods that rank order examinees in a consistent manner at any given time, but that show less consistency in ranking examinees across different times. This phenomenon is of potential interest in measuring progress in language acquisition, since it may serve to identify skill areas, language features, and task formats that are more or less consistently learned by members of the examinee population. Those skills, format types, and language features that can be shown to exhibit greater regularity in rank ordering learners from time 1 to time 2 would likely be superior candidates for inclusion in a language test battery when it is suspected that learner- and situation-specific errors may threaten measurement accuracy. Identification of differential acquisition and attrition rates for selected features of language continues to be an area of useful empirical study (Brown, 1972, 1987).

Related to the issue of appropriate reliability estimation is the consideration of "item bounce," or the stability of item statistics over time. It is useful to determine which item formats exhibit the highest proportion of items with stable characteristics. This information would be appropriate to gather if a certain component or subset of items of the test were to be selected for anchoring purposes in the equating process. If, for example, the multiple-choice, written-expression component were used for anchoring in the equating of TWE essay prompts, it would be useful to determine the content and format characteristics of items with the greatest comparative stability of item difficulty or discriminability for that component. It is expected that items involving some language features, component skills, or specialized formats will have more stable difficulty estimates than items involving other language features, component skills, or specialized formats.

Although eventually such a study might also be useful for the Test of Spoken English (TSE®), the Test of Written English (TWE®), and other TOEFL Program examinations, for the present study it was intended that only the TOEFL itself would be so analyzed. Based on results of such an initial study, it was hoped that similar studies might be conducted with other examinations at some future time.

Purpose

The present study was proposed to provide test-retest reliability estimates for the TOEFL total and component scores. The study also reported internal-consistency and alternate-form reliability estimates for those same total and component scores as comparative measures. Additionally, the study intended to carefully consider the subtest statistical characteristics within components and formats across the two times of administration. In this way not only would global reliability estimation be provided of the kind normally reported in examination user manuals, but also reliability and stability information would be made available within components and format groupings as

an indication of the comparative consistencies of various parts of the TOEFL test battery. After control was made for the comparative number of items within component and format type, it was intended that this information could be useful as a further indication of format quality for purposes of test development and test equating. It was hoped that insights also might be gained concerning the nature of language competence and processes of language acquisition by attending to within-skill performance changes on repeated assessment.

Method

Subjects

Exactly 329 subjects participated by individual informed consent from among the ESL students enrolled in the large intensive English programs at three state-owned universities in Southern California. The project proposal called for testing approximately 300 subjects, with an additional 10 percent to be tested to compensate for possible attrition. These university testing sites were selected on the following bases:

- availability of large numbers of students
- rough approximation of student characteristics to the known characteristics of the global TOEFL examinee population
- participation or potential participation of the programs in the Institutional TOEFL Program
- willingness to participate in the research project

Of the 329 subjects who participated, 60 appeared only at the first or last test administration, so repeated measures were available for only 269 subjects. The precise distribution of subjects by language group and repetition condition is reported in Table 1. Note that the predominantly Asian language distribution of the subjects was not dissimilar to that of the current TOEFL examinee population, although the proportion of Far Eastern students in the present study may have been somewhat higher than usual TOEFL administrations (i.e., roughly 70 percent *vs.* the usual 50 percent).

Full confidential disclosure of TOEFL scores was provided at no charge to subjects after the second administration as partial incentive for participation. In practice this meant that students received scores from the higher of the two administrations delivered confidentially on unofficial institutional report forms. This TOEFL practice opportunity provided sufficient incentive for those subjects who participated, since they were primarily studying English with the expressed purpose of improving TOEFL scores to enhance their eligibility for university admission.

Table 1

Sample Description by Language Background
for Repeating and Non-Repeating Subjects

Language	Repeating[1]	Non-Repeating[2]	Total
Arabic	8	1	9
Chinese	75	16	91
French	4	2	6
German	3	0	3
Indonesian	13	1	14
Italian	7	2	9
Japanese	47	7	54
Korean	61	21	82
Persian	4	0	4
Portuguese	4	3	7
Russian	1	0	1
Spanish	7	2	9
Tagalog	3	1	4
Thai	26	3	29
Turkish	6	1	7
	269	60	329

[1] Repeating subjects are those for whom retest data was available because they appeared for testing at both administrations.

[2] Non-repeating subjects are those for whom retest data was not available because they appeared for testing at only one test administration.

Instrumentation

Two disclosed forms of the TOEFL test from the July and August 1991 administrations were used. These two forms are hereafter labeled "Forms A and B" respectively. The advantages of these forms were that they represented current TOEFL test format and there was reason to believe that the students tested were unlikely to have encountered either form at any time in the past.

Procedures and Design

The research design called for the two disclosed forms of the TOEFL test, A and B, to be administered to four randomly determined subgroups of the experimental sample according to the pattern in Table 2, below.

Table 2

The Design for Repeated Administrations of TOEFL Forms A and B to Randomly Selected Subgroups of the Sample

	N (Actual)[1]	Test	Retest
Group A	100 (101)	TOEFL A	TOEFL A
Group B	100 (91)	TOEFL B	TOEFL B
Group C	50 (25)	TOEFL A	TOEFL B
Group D	50 (52)	TOEFL B	TOEFL A

[1] Numbers outside parentheses indicate planned group sizes. Numbers in parentheses indicate actual numbers of subjects who appeared, after elimination of 60 subjects who failed to attend either the first or second administration.

At time 1, subjects were assigned by random stratification on course level to groups within site, to comprise a total approximating and slightly exceeding the planned total within cells. A minimum of two large rooms were reserved at each site to permit simultaneous testing of TOEFL Forms A and B. Because 60 subjects, as indicated in Table 1, did not attend both testing sessions at their respective institutions (the majority of the 60 came at time 1 and not at time 2, but several came at time 2 and not at time 1), the numbers of repeating subjects in Groups B and C were somewhat less than originally planned. As Table 2 indicates, the time 2 attrition was greatest for Groups B (91 repeaters) and C (25 repeaters). In particular, due to scheduling and other difficulties, approximately 25 subjects tested at time 1 in Group C were not available for time 2 testing. The majority of those not appearing at time 2 for Group C were unavailable for unanticipated reasons. It therefore cannot be affirmed that for those who returned at time 2, some systematic pattern of attrition did not alter the random nature of the group established at time 1. At each site, a list of subjects was compiled in advance, consisting of the names and English course levels of those who agreed to participate. The participants were assigned to design groups at time 1 by random stratification on course level to ensure a roughly proportional representation by course level in each group. At time 2, the same group assignments were maintained as at time 1, regardless of attrition patterns. Although the ultimate size of Group C was disappointingly small, subjects were tested in sufficient numbers to permit the planned analyses of response data.

Rationale for the Length of the Intervening Time Period

The interval between test and retest for each participant was planned as less than one week and no more than two weeks, to be determined exactly on the basis of program considerations. In practice, the interval was exactly eight days for every subject. Subjects were not provided with performance feedback until after the second administration. This absence of feedback was intended to minimize the potential effects of further directed learning exposures and to permit subjects to forget much of the content of the first test. The presence and magnitude of practice or exposure effects was intended to be partially apparent by comparing the performance of Groups C and D (by computation of means and standard deviations and alternate-form reliability estimates) with the performances of Groups A and B (by computation of means and standard deviations and test-retest reliability estimates). Also, internal-consistency estimates were provided for the TOEFL test in all forms, components, testing times, and examinee groups. Given the length of each test form (146 scored items), the eight-day interval between administrations was considered a conventionally appropriate period to permit the subjects to forget content from their initial exposure, so as to reduce practice effects while not allowing time for the examinees to acquire substantive maturational and language acquisitional changes (Henning, 1987).

If more time had been permitted in the interval, there would have been greater likelihood that changes would have taken place. Allowing for language learning or forgetting by increasing the time interval might have been interesting from the viewpoint of the study of language acquisition; however, the focus of this study was the stability of test and subtest functioning

rather than the nature of the language acquisition process. Also, since there was no control over the learning exposure available during the interval, it was recognized that a prolonged interval would allow different kinds and amounts of English language learning or forgetting on the part of different examinees, a phenomenon that could unfairly affect the test-retest reliability estimates obtained.

In this connection, while Hale, Angelis, & Thibodeau (1980) report that there was a small but significant effect on TOEFL performance due to prior disclosure (4.6 percent average improvement), in their study examinees were universally provided with disclosed forms and encouraged to review them over a period of several weeks to prepare for subsequent examinations, some of which employed the disclosed items. This study was purposely different in several ways. First, examinees were not given test forms or feedback on initial testing to prepare for subsequent administrations. Second, the interval between pretest and posttest was shorter here than in the Hale et al. study. Third, the primary focus of research in the present study was on comparative reliability and stability of subtest item types across administrations, rather than on comparison of mean exposure effects. Thus, learning exposure in the present study was purposely minimized.

Analyses

Descriptive statistics were computed for all forms, testing times, and examinee groups. Reliability estimates were made for all forms, components, testing times, and examinee groups, including internal-consistency, test-retest, and alternate-form reliability estimates. The Spearman-Brown Prophecy Formula was used to permit the comparative study of adjusted reliabilities with item number held constant across subtests. It is recognized that such adjustments assume that the test length would be expanded by adding new test items drawn from a domain of test items included only in parallel test forms (Gulliksen, 1987). While it is most common to apply Spearman-Brown adjustments in situations involving internal consistency reliability estimation, Gulliksen (p. 215) affirms that it is appropriate for other kinds of reliability estimation also, provided that the criterion of parallel forms can be satisfied. Therefore, test-retest applications of Spearman-Brown adjustments could be appropriate in situations where there is no significant difference between variances or covariances at the two testing times.

Item types were studied within subtests by means of subtest means, standard deviations, and reliabilities. In this way it was thought that subtests consisting of items with comparatively less stable difficulty values would become evident (i.e., comparisons were made across administrations to identify item types that exhibit "item bounce" phenomena).

Results

Mean Differences Across Administration Times

Table 3 reports the means, standard deviations, and matched-group *t* values for TOEFL Form A for 101 repeating examinees across two test administrations over an eight-day interval. It is of interest to determine whether there was improvement in performance on the same test in the second trial. More particularly, it is important to ascertain whether any observed changes in performance were statistically significant or were specific to any particular test item type. It should be noted here that operational administrations of the TOEFL test specifically exclude the possibility of repeated administration of the same test to the same examinees. Thus, the information provided by investigating mean differences over time with the same examinees and the same test forms is of interest primarily for research applications and to gain a better understanding of differential stabilities of subtest difficulties to test development, but does not represent a real or expected operational situation.

In general, results in Table 3 suggest, predictably, that performance change was primarily in the direction of score improvement at time 2. The improvement was most pronounced in all three varieties of listening comprehension items (i.e., multiple-choice statements, dialogues, and minitalks) and in the reading comprehension subtest. While the author can point to no particular theory to account for this differential change, it has been observed elsewhere that listening skill appears to be among those language skills that are most rapidly acquired and most rapidly forgotten, depending upon the learning environment (Henning, 1982). Also, the listening and reading item types rely more heavily on context to provide the semantic information necessary for successful responses than the other item types in the TOEFL test.

The observed differential improvement phenomenon calls to mind a fundamental distinction between language learning and language assessment. From the perspective of test development and accurate assessment, it would seem that other things being equal, among the most desirable item types to include in large-scale testing might be those that are least susceptible to performance improvement attributable to prior exposure to the test items. By contrast, from the perspective of language teaching, those elements of a language activity that are most likely to promote learning are usually considered most desirable to include in an assessment. In the present example, for TOEFL Form A it would appear that the sections of the test that showed least improvement over time were the structure and written expression section and the vocabulary section. It does not appear that this lack of performance improvement for those item types was at all due to ceiling effects, since the means for all item types appear to be near the mid-points of the possible scoring distributions. It would be important to replicate this tendency over other forms of the test, as was done next, before making any subtest-specific generalizations. It is also important to determine whether any performance increment is uniform across subjects, or whether it is attached to examinees differentially, thus changing examinee rank order on retesting. This latter concern is addressed later.

Table 3

Means, Standard Deviations, and Matched-Group t Values for TOEFL Form A Subtest and Total Score Same-Sample Comparisons at Time 1 and Time 2 (N = 101)

		Time 1 (Form A)		Time 2 (Form A)		
Subtest	N Items	Mean	SD	Mean	SD	t
Listening	50	25.10	8.71	28.29	9.47	6.42**
Statements	20	10.08	3.91	11.20	4.13	3.85**
Dialogues	15	7.06	3.34	8.25	3.53	5.23**
Minitalks	15	7.96	2.79	8.84	2.91	3.81**
Structure & Written Expression	38	22.24	7.54	22.28	7.82	0.09
Structure	14	8.58	2.85	8.59	3.04	0.04
Written Expression	24	13.65	5.30	13.68	5.45	0.09
Vocabulary & Reading	58	32.76	10.39	33.76	11.02	1.31
Vocabulary	29	16.04	5.78	15.93	6.51	-0.21
Reading	29	16.72	5.33	17.83	5.81	2.74**
TOEFL Total (Raw Score)	146	80.10	24.04	84.33	25.52	3.29**
Listening (Scaled Score)	50	46.69	5.69	48.84	6.10	6.42**
Structure & Written Expression (Scaled Score)	38	45.86	8.19	45.85	8.58	-0.02
Vocabulary & Reading (Scaled Score)	58	44.81	7.28	45.49	7.74	1.23
TOEFL Total (Scaled Score)	146	457.90	63.61	467.25	67.44	2.53*

*p < 0.05
**p < 0.01

Table 4

Means, Standard Deviations, and Matched-Group t Values for TOEFL Form B Subtest and Total Score Same-Sample Comparisons at Time 1 and Time 2 (N = 91)

		Time 1 (Form B)		Time 2 (Form B)		
Subtest	N Items	Mean	SD	Mean	SD	t
Listening	50	26.69	9.10	29.44	9.40	6.31*
Statements	20	10.42	4.37	11.47	4.62	3.49*
Dialogues	15	7.78	2.96	8.36	3.10	2.73*
Minitalks	15	8.50	2.87	9.60	2.75	6.10*
Structure & Written Expression	38	20.08	7.57	21.71	7.30	2.80*
Structure	14	7.76	3.50	8.28	3.25	1.71
Written Expression	24	12.32	4.60	13.44	4.63	2.88*
Vocabulary & Reading	58	33.36	10.53	36.82	9.67	4.68*
Vocabulary	29	15.98	5.64	17.07	5.39	2.84*
Reading	29	17.39	5.74	19.76	4.88	4.88*
TOEFL Total (Raw Score)	146	80.13	23.31	87.98	24.31	6.41*
Listening (Scaled Score)	50	48.41	5.81	50.08	6.06	5.56*
Structure & Written Expression (Scaled Score)	38	44.86	7.96	46.79	6.85	2.83*
Vocabulary & Reading (Scaled Score)	58	45.47	7.43	47.85	6.71	4.46*
TOEFL Total (Scaled Score)	146	462.47	59.07	482.39	59.87	5.49*

*$p < 0.01$

Accordingly, Table 4 reports on a replication of the Form A results in Table 3, except that Table 4 includes results from the application of Form B with a randomly different subsample of 91 subjects from the same three institutions considered in Table 3. Note that while performance improvement was more evident and generalized across subtests for Form B than for Form A, again there was an observable tendency for listening and reading comprehension subtests to show more performance improvement and less stability over administrations than was the case for structure and vocabulary subtests. It would seem that this tendency of differential improvement in score was consistent across the two administrations with two different versions of the test.

Since this pattern of differential improvement appeared consistent across test forms and administrations, it is useful to consider item type characteristics that may have contributed to this result. One possible explanation for the particular pattern of differential improvement is that the item types showing performance improvement all involved comprehension of meaning from context, while the item types that did not show performance improvement were less dependent on context for meaning. With regard to the vocabulary subtest, for example, it has been observed elsewhere (Henning, 1991) that even though the TOEFL vocabulary section presents words in context, there is generally and intentionally little information provided in the vocabulary item stems that could be used to infer the meaning required to choose the correct option. Consistent with this explanation, vocabulary was one of those sections of the test that showed least improvement on retesting after an eight-day interval. Although some subtest and total score means differed significantly over time as noted, Tables 3 and 4 indicate that there were no observable patterns of change in the variances of scores by subtests over time.

Mean Differences Across Test Forms

It is also important to consider whether the observed mean score improvement over time was due to the participants' ability to remember information from the first administration to the second, or whether the improvement was attributable more to practice and familiarity with the particular item types and testing procedures. Tables 5 and 6 report differences on repeated measures when the test form was changed from time 1 to time 2. Table 5 reports means, standard deviations, and matched-group *t* values for 25 examinees who responded to Form A first and to Form B eight days later. Table 6 reports the same information for 52 other examinees who encountered Form B first, followed by Form A. Note from Table 5 that there was a general tendency for score improvement on Form B over Form A. Some slight improvement on the second administration might be predicted on the basis of increased familiarity with testing format and procedures, even when different versions of the test are employed. In this case, however, performance differences may not have been due entirely to practice and familiarity effects, but also to an overall tendency for Form B content to be

Table 5

Means, Standard Deviations, and Matched-Group t Values for Subjects Administered TOEFL Form A Before TOEFL Form B (N = 25)

	Time 1 (Form A)			Time 2 (Form B)		
Subtest	N Items	Mean	SD	Mean	SD	t
Listening	50	23.84	7.58	26.40	8.09	1.78
Statements	20	10.08	3.27	10.04	4.11	-0.05
Dialogues	15	6.76	3.13	8.00	3.14	2.64*
Minitalks	15	7.00	2.35	8.36	2.23	2.76*
Structure & Written Expression	38	20.80	6.44	22.00	6.14	1.59
Structure	14	8.32	2.36	8.12	2.49	-0.56
Written Expression	24	12.48	4.46	13.88	4.11	2.26*
Vocabulary & Reading	58	32.96	10.33	35.88	8.35	2.44*
Vocabulary	29	16.64	6.09	17.20	4.96	0.70
Reading	29	16.32	4.97	18.68	4.22	3.30**
TOEFL Total (Raw Score)	146	77.60	20.73	84.24	19.39	2.49*
Listening (Scaled Score)	50	45.80	5.21	48.04	5.12	2.20*
Structure & Written Expression (Scaled Score)	38	44.40	6.69	46.96	5.37	3.22**
Vocabulary & Reading (Scaled Score)	58	44.80	7.11	47.16	5.59	2.70*
TOEFL Total (Scaled Score)	146	449.96	54.34	473.88	45.86	3.46**

*p < 0.05
**p < 0.01

Table 6

Means, Standard Deviations, and Matched-Group t Values for Subjects Administered TOEFL Form B Before TOEFL Form A (N = 52)

		Time 1 (Form B)		Time 2 (Form A)		
Subtest	N Items	Mean	SD	Mean	SD	t
Listening	50	22.64	8.80	23.79	7.41	1.53
Statements	20	8.44	3.77	9.54	3.37	2.53*
Dialogues	15	6.83	3.30	6.56	2.58	-0.84
Minitalks	15	7.37	2.83	7.69	2.62	0.81
Structure & Written Expression	38	19.12	7.52	20.25	6.99	1.51
Structure	14	7.37	3.65	8.12	2.75	1.69
Written Expression	24	11.75	4.68	12.14	4.86	0.67
Vocabulary & Reading	58	32.42	9.78	29.10	9.55	-3.42**
Vocabulary	29	15.56	5.50	14.36	5.38	-2.06*
Reading	29	16.87	5.20	14.73	5.13	-3.76**
TOEFL Total (Raw Score)	146	74.17	23.23	73.14	21.05	-0.58
Listening (Scaled Score)	50	45.42	6.09	45.96	4.70	0.98
Structure & Written Expression (Scaled Score)	38	43.62	7.64	44.12	7.28	0.56
Vocabulary & Reading (Scaled Score)	58	44.77	6.67	42.17	6.75	-3.73**
TOEFL Total (Scaled Score)	146	446.06	59.54	440.85	54.75	-0.97

*p < 0.05
**p < 0.01

easier than Form A content for these examinees. Inspection of Table 6 reveals that the significant subtest mean differences over time were in the reverse order for the vocabulary and reading sections; there appeared to be a performance decline at time 2 with Form A for that particular examinee group. These combined results reported in Tables 5 and 6 support the view that Form B was easier than Form A, especially with regard to the vocabulary and reading component. Alternative explanations of this outcome are also possible, including the possibility that motivational or administrative differences caused the group responding to Form A at time 2 to perform below expectation. This alternative explanation is not entirely satisfying, since this group was comprised of persons tested at three separate testing sites who were usually in the same room as others who were taking the same test form on the same day. This finding of group performance differences on the vocabulary and reading comprehension components of Forms A and B, however, does not affect earlier generalizations based on information in Tables 3 and 4, where comparisons were made across two administrations of the same test version.

Note also that the possible tendency for Form B to be easier than Form A, at least with regard to the vocabulary and reading components, appeared to persist across administrations in terms of both scaled and raw scores reported. Technically speaking, tests like TOEFL exist in equated forms that permit comparable inferences about scores across versions, even though such tests do not necessarily exist in equivalent forms with equal means, variances, and covariances across all test versions or forms. Therefore, it is not surprising that there would be mean differences across forms in observed raw scores. Normally, however, one would expect difficulty differences observed in the raw scores to disappear in the scaled scores because of the equating procedures involved in deriving the scaled scores. Admittedly this part of the study involved use of small subject samples, with only 25 and 52 subjects for the two administrations. This is a much smaller and less representative sample than employed in the actual equating performed with each new TOEFL form. It is also possible that alternative explanations such as those mentioned above could account for the differential performance in vocabulary and reading across the two test forms. Nevertheless, these possible findings with regard to the small but significant changes in scaled scores across two TOEFL forms suggest the value of a larger empirical study of the accuracy of the TOEFL equating procedure across various TOEFL test forms and administrative time intervals.

Note that while there was a slight tendency for score variance to decrease from time 1 to time 2 in the listening comprehension and vocabulary and reading comprehension components for both Forms A and B, there was no such consistent pattern for the structure and written expression component, where the score variance appeared to increase for Form A and decrease for Form B. These patterns of change, however, did not attain statistical significance, apart from the decrease in Form B variance from time 1 to time 2 for structure and written expression ($F_{max} = 2.02$; $p < 0.05$; $df = 24, 51$). As with changes in subtest means reported in Tables 2 and 3, changes in subtest variance did not appear to be attributable to any ceiling effects.

Test-Retest and Internal-Consistency Reliability Estimates

Although it is important to note patterns of score change over time as reported in Tables 3 through 6, it should be recognized that average score differences, even if statistically significant, may not affect the rank ordering of examinees within or across testing times. If practice or learning effects are uniformly and equally distributed across subjects, then rank order does not change and the ability of the test to differentiate among examinees' proficiency is not affected. Since some mean differences over time were observed in Tables 3 and 4, it is all the more important that evidence of test-retest reliability be available to indicate that examinee ranking is not significantly altered on repeated test administrations within controlled, short intervals of time. Tables 7 and 8 provide both test-retest and internal consistency reliability estimates for TOEFL Forms A and B on samples of 101 subjects (in the case of Form A) and 91 subjects (in the case of Form B).

Noting the first two columns of coefficients in Table 7, we see that test-retest reliabilities were all respectably high for Form A. Because of the known relationship between test length in number of items and reliability, which is that shorter tests are usually less reliable than longer tests, the Spearman-Brown Prophecy Formula was used to project expected reliabilities if test lengths were equivalent. The applications and assumptions of this procedure are explained by Gulliksen (1987). For this comparison, the same projected 146-item length was arbitrarily used, as was the actual length of the total TOEFL test. We can see in the second column of coefficients that when test length was adjusted to be constant in this way, all of the coefficients, except the raw scores of the vocabulary & reading comprehension component, exceeded .90; that exception exhibited adjusted test-retest reliability of .88. This same pattern of acceptably high reliabilities persisted with Form B, as reported in Table 8. Again, the lowest adjusted subtest reliability was that for the raw scores of the vocabulary & reading comprehension component, which, in the case of Form B, was .89. It is important to point out, however, that the reliabilities of scaled scores for the vocabulary & reading comprehension component, which are the scores actually reported, were low only in the case of Form B (viz., .88). The comparatively lower magnitudes of test-retest reliability estimates for raw scores of the vocabulary & reading comprehension component that were observed also do not reflect statistically significant differences for all possible comparisons, even though the same pattern was apparent in the case of the raw scores of both Form A and Form B.

With regard to a comparison between test-retest and internal-consistency reliability estimates, it is noteworthy that across forms, the listening comprehension component consistently tended to exhibit higher test-retest than internal-consistency reliability. For the other components of the test, the differences between test-retest and internal-consistency reliability estimates were consistently in the opposite direction.

Among the listening subsections of the TOEFL test, the minitalks section exhibited the lowest internal consistency reliability for both Forms A and B. Although the adjusted coefficients would not be significantly different from

Table 7

Test-Retest and Internal-Consistency (KR21) Reliability Coefficients Adjusted and Unadjusted for 146-Item Test Length with TOEFL Form A Repeated (N = 101)

		Times 1 & 2 (Form A)		Time 1 (Form A)		Time 2 (Form A)	
Subtest	N Items	R_{AA}	$R_{AA}(146)$	KR21	KR21(146)	KR21	KR21(146)
Listening	50	.85	.94	.85	.94	.88	.96
Statements	20	.74	.95	.71	.95	.75	.96
Dialogues	15	.78	.97	.71	.96	.75	.97
Minitalks	15	.67	.95	.56	.93	.61	.94
Structure & Written Expression	38	.82	.95	.86	.96	.87	.96
Structure	14	.66	.95	.64	.95	.69	.96
Written Expression	24	.80	.96	.83	.97	.84	.97
Vocabulary & Reading	58	.74	.88	.88	.95	.90	.96
Vocabulary	29	.63	.90	.81	.96	.86	.97
Reading	29	.74	.93	.78	.95	.83	.96
TOEFL Total (Raw)	146	.87	.87	.94	.94	.95	.95
Listening (Scaled)	50	.84	.94	---	---	---	---
Structure & Written Expression (Scaled)	38	.75	.92	---	---	---	---
Vocabulary & Reading (Scaled)	58	.88	.95	---	---	---	---
TOEFL Total (Scaled)	146	.85	.85	---	---	---	---

$p < 0.05$ - One-Tailed Critical Value = 0.164
$P < 0.01$ - One-Tailed Critical Value = 0.230

Table 8

Test-Retest and Internal-Consistency (KR21) Reliability Coefficients Adjusted and Unadjusted for 146-Item Test Length with TOEFL Form B Repeated (N = 91)

		Times 1 & 2 (Form B)		Time 1 (Form B)		Time 2 (Form B)	
Subtest	N Items	R_{BB}	$R_{BB}(146)$	KR21	KR21(146)	KR21	KR21(146)
Listening	50	.90	.96	.87	.95	.88	.96
Statements	20	.80	.97	.78	.96	.81	.97
Dialogues	15	.78	.97	.61	.94	.66	.95
Minitalks	15	.81	.98	.59	.93	.58	.93
Structure & Written Expression	38	.72	.91	.86	.96	.85	.96
Structure	14	.64	.95	.77	.97	.73	.97
Written Expression	24	.68	.93	.75	.95	.76	.95
Vocabulary & Reading	58	.76	.89	.89	.95	.87	.95
Vocabulary	29	.78	.95	.80	.95	.79	.95
Reading	29	.63	.90	.82	.96	.76	.94
TOEFL Total (Raw)	146	.88	.88	.94	.94	.95	.95
Listening (Scaled)	50	.88	.96	---	---	---	---
Structure & Written Expression (Scaled)	38	.62	.86	---	---	---	---
Vocabulary & Reading (Scaled)	58	.75	.88	---	---	---	---
TOEFL Total (Scaled)	146	.83	.83	---	---	---	---

$p < 0.05$ - One-Tailed Critical Value = 0.173
$p < 0.01$ - One-Tailed Critical Value = 0.242

the coefficients for other subsections of the test, the pattern is replicated across forms. It is mentioned here also since this finding of comparatively lower internal consistency for listening minitalks has been found in previous studies by the author (Henning, 1991). Interestingly, this regularly lower internal consistency for listening minitalks appeared to be partially offset by comparatively higher test-retest reliability for that section, especially with regard to Form B, as reported in Table 8. If there were any noticeable patterns of difference between test-retest and internal-consistency reliability estimates, they would probably only be the observation that minitalks appeared to show slightly higher test-retest than internal-consistency reliability estimates, whereas estimates for the vocabulary and reading comprehension component tended to differ in the opposite direction.

Internal-consistency estimates in the present study were obtained by use of Kuder-Richardson Formula 21 (Gulliksen, 1987). This was done for convenience and because it was known that in many standard applications this procedure provides a slightly lower estimate of the actual reliability than Kuder-Richardson Formula 20 or Cronbach's alpha, so that generalizations would usually be conservative. Sireci, Thissen, & Wainer (1991) point out that there is a tendency for K-R 20 and alpha estimates to be inflated for context dependent testlets if there is a violation of the constraints of local independence. Also, since these internal-consistency estimates were developed as a means of estimating alternate-form reliability without repeated testing, the preferred internal-consistency procedure should be the one that corresponds most closely to the correlation between alternate forms, which is also available in this study. In the case of this particular study, the K-R 21 procedure provided better approximation of alternate-form reliability estimates than the K-R 20 procedure.

Alternate-Form Reliability Estimates

Tables 9 and 10 report the correlations between scores for Forms A and B for test subsections and totals. Table 9 reports these correlations when Form A was administered before Form B with a sample of 25 examinees. Table 10 reports the same kind of information when Form B was administered before Form A with a sample of 52 examinees. Again, the Spearman-Brown Prophecy Formula was used to provide reliability estimates for subtests with uniform 146-item test length. In the first column of coefficients reported in Tables 9 and 10, it appears that correlations ranged from a low of .31 to a high of .83. After the Spearman-Brown adjustments to hold test length constant, the correlations in the second column of those tables ranged from a low of .77 for listening statements to a high of .97 for structure. Because these extreme estimates issued from the performance of a group of only 25 subjects, they are not particularly stable estimates and should not cause the same concern or satisfaction as if they had been replicated exactly with the larger group of 51 subjects. Of course, a similar limitation is present for both the lower estimate, .77, and the higher estimate, .97, because they are similarly based on small sample size.

Table 9

Alternate-Form and Internal-Consistency (KR21) Reliability Coefficients Adjusted and Unadjusted for 146-Item Test Length with TOEFL Form A Administered at Time 1 and TOEFL Form B Administered at Time 2 (N = 25)

		Times 1 & 2 (Form A, B)		Time 1 (Form A)		Time 2 (Form B)	
Subtest	N Items	R_{AB}	R_{AB}(146)	KR21	KR21(146)	KR21	KR21(146)
Listening	50	.58	.80	.80	.92	.83	.93
Statements	20	.31	.77	.56	.90	.74	.95
Dialogues	15	.72	.96	.66	.95	.66	.95
Minitalks	15	.42	.88	.34	.	.28	.79
Structure & Written Expression	38	.82	.95	.79	.94	.78	.93
Structure	14	.73	.97	.42	.88	.48	.91
Written Expression	24	.74	.95	.73	.94	.68	.93
Vocabulary & Reading	58	.82	.92	.88	.95	.82	.92
Vocabulary	29	.76	.94	.84	.96	.74	.94
Reading	29	.71	.92	.73	.93	.65	.90
TOEFL Total (Raw)	146	.78	.78	.92	.92	.91	.91
Listening (Scaled)	50	.52	.76	---	---	---	---
Structure & Written Expression (Scaled)	38	.80	.94	---	---	---	---
Vocabulary & Reading (Scaled)	58	.79	.90	---	---	---	---
TOEFL Total (Scaled)	146	.77	.77	---	---	---	---

$p < 0.05$ - One-Tailed Critical Value = 0.330
$p < 0.01$ - One-Tailed Critical Value = 0.454

Table 10

Alternate-Form and Internal-Consistency (KR21) Reliability Coefficients Adjusted and Unadjusted for 146-Item Test Length with TOEFL Form B Administered at Time 1 and TOEFL Form A Administered at Time 2 (N = 52)

		Times 1 & 2 (Form B, A)		Time 1 (Form B)		Time 2 (Form A)	
Subtest	N Items	R_{BA}	R_{BA}(146)	KR21	KR21(146)	KR21	KR21(146)
Listening	50	.60	.81	.86	.95	.79	.92
Statements	20	.74	.95	.69	.94	.59	.91
Dialogues	15	.62	.94	.71	.96	.48	.90
Minitalks	15	.67	.95	.57	.93	.49	.90
Structure & Written Expression	38	.72	.91	.85	.96	.83	.95
Structure	14	.53	.92	.79	.98	.59	.94
Written Expression	24	.62	.91	.76	.95	.78	.96
Vocabulary & Reading	58	.74	.88	.87	.94	.85	.93
Vocabulary	29	.71	.92	.79	.95	.78	.94
Reading	29	.69	.92	.77	.93	.75	.93
TOEFL Total (Raw)	146	.83	.83	.94	.94	.92	.92
Listening (Scaled)	50	.76	.90	---	---	---	---
Structure & Written Expression (Scaled)	38	.63	.87	---	---	---	---
Vocabulary & Reading (Scaled)	58	.72	.87	---	---	---	---
TOEFL Total (Scaled)	146	.78	.78	---	---	---	---

$p < 0.05$ - Critical Value = 0.231
$p < 0.01$ - Critical Value = 0.322

Guilford & Fruchter (1973) note that the significance of the difference of a correlation coefficient from zero can be estimated as a t ratio, assuming a bivariate normal distribution in the population. For a sample of 25 persons, the worst-case scenario in the present study, a correlation of about .33 would be required to achieve statistical significance above zero at the $p < 0.05$ level. By reference to the unadjusted coefficients in column 2 of Tables 8 and 9, it is apparent that all but the alternate forms correlation for listening statements in the 25-person sample situation (viz., .31) satisfied this criterion. With regard to a consideration of the magnitude of difference between correlation coefficients that could be called significant, Guilford & Fruchter present methods for comparing coefficients derived using the same variables with different and unmatched samples, and again using the same criterion variable for different compared variables with the same sample of persons. In the present example, a comparison of subtest correlations would consist of a comparison of coefficients using different variables with the same sample of persons.

Conclusions

The present study was conducted to derive comparative estimates of test-retest, alternate forms, and internal-consistency reliability across TOEFL forms and subsection item types. In addition, the study considered patterns of subsection difficulty change over one eight-day interval with no intervening feedback on performance.

Admittedly, there were numerous limitations of the present study, as noted earlier. Sample size was relatively small and not perfectly representative of the usual TOEFL population, whether in terms of language background or overall mean proficiency. Estimates of reliability are known to be highly sample-dependent, so that sampling constraints are important. The study employed only two TOEFL forms, so it is not clear how accurately the findings can be generalized to other TOEFL forms. Many of the differences noted among coefficients for particular subtests did not attain statistical significance. Nevertheless, there was a tendency for patterns of results to be replicated across forms and samples, as discussed.

The most salient findings for the particular test forms and examinee samples employed in the study may be summarized as follows:

1. Test-retest reliability estimates were adequately high across reported components and total test raw scores for the two forms investigated. Test-length-adjusted coefficients ranged from .87 to .98 in magnitude, with a mean coefficient of .93, over 22 total observed raw-score test-retest reliability estimates.

2. There was a replicated pattern of comparative test-retest reliability estimates, with the vocabulary and reading component exhibiting the lowest test-retest reliability across both forms of the test (viz., .88 and .89). This finding must be qualified in two ways: first, the pattern persisted only in the case of raw-score comparisons and not scaled-score comparisons; second,

the difference between these adjusted coefficients and the nearest comparison coefficients did not reach statistical significance. This finding was possibly related also to the pattern of significant score improvement on retesting noted for passage-dependent items, as described in number 6, below. Since the vocabulary & reading component contained both context-dependent reading-inference items and non-context-dependent vocabulary-knowledge items, the unique combinations of these items may have led to different rank orderings of examinees on proficiency over repeated administrations of the same test forms. Although this speculation may be interesting, it is noted again that the test-retest reliability estimates for vocabulary & reading were not significantly lower than test-retest reliability estimates for other subsections of the test.

3. Alternate-form reliability estimates were consistently lower than either test-retest or internal-consistency reliability estimates, with test-length adjusted coefficients ranging from .77 to .97 in magnitude. The mean reliability by this method was .90, averaged over 22 raw-score alternate-form reliability estimates. The extreme low and high estimates (viz., .77 and .97) may be partially attributable to the small sample size of one of the groups investigated (viz., 25 persons). In a group of 52 persons employed for alternate-form reliability estimation, the extreme estimates were only .81 and .95 for the same tests.

4. Internal consistency reliability estimates, derived by the Kuder-Richardson Formula 21 procedure, ranged in magnitude from .79 to .98 when adjusted to uniform 146-item test length by means of the Spearman-Brown Prophecy Formula. The mean internal-consistency estimate was .94 when averaged over 88 separate raw-score coefficients. The lowest estimate, .79 for listening minitalks, may be partially explained by noting that the group size was small for that estimation, consisting of only 25 persons, and the K-R 21 procedure is known to underestimate internal consistency by comparison with other estimation methods under certain circumstances.

5. There was a patterned tendency for the listening minitalks subsection of the test to exhibit the lowest internal consistency of all parts of the test across both forms and both testing times, with adjusted estimates ranging from .79 to .93. This tendency seemed to be offset by the tendency for that same subsection to exhibit comparatively high adjusted test-retest reliability, .95 and .98, for the two forms of the test, A and B respectively. This disparity between reliabilities estimated for Listening Minitalks by the two different methods may be explained in the following manner. Internal consistency reliability provides an indication of the homogeneity of item variances within a given test form. Test-retest reliability, in contrast, reflects the tendency of examinees to perform in the same rank order of attainment on separate administrations of the same test, regardless of how homogeneous the item content within the test. It is known that the Listening Minitalks subsection of the test has exhibited comparatively high variation in the numbers and kinds of listening passages from one version of the test to the next, including extended dialogs and brief lectures. In short, the items in this subsection have tended to be less homogeneous in content and format than the items in other sections of the test, which would account for the comparatively lower internal consistency of this subsection. The

comparatively high test-retest reliability estimates for the same subsection suggest that the comparative lack of item homogeneity did not appear to affect the rank ordering of examinees on repeated application of the same tests.

6. There was a pattern of significant score improvement from time 1 to time 2 for both Form A and Form B of the test. However, the pattern was such that the improvement was consistent only in the listening and reading comprehension portions of the test. Subtests of structure and vocabulary were less consistent, or failed to show the same level of score improvement over time. It was speculated that those portions of the test that depended less on the ability to infer meaning from context to arrive at correct answers also were less susceptible to practice effects over repeated administrations of a test form. In the case of the TOEFL test, where the practice has been not to permit the repeated use of the same test form with the same examinees, the possibility that context-dependent items may be more susceptible to practice effects than non-context-dependent items is not considered a threat to reliability in operational administrations.

The question of statistical significance of coefficients and of differences among coefficients bears further comment at this point. It is not possible to set a single statistical criterion for comparisons among these coefficients because the coefficients were generated using four individual sample sizes ranging from 25 to 101, length-adjusted raw-score reliability coefficients alone ranged in magnitude from .77 to .98 over 132 computed coefficients, and both matched-group and unmatched-group comparisons would be entailed. It was noted earlier in this report that we can consider, for example, the worst-case scenario, where the reliability estimates were generated using the 25-person group. In this case, correlations not adjusted for length would need only to exceed the critical value of .33 in order to attain one-tailed statistical significance above zero at the $p < .05$ level. For larger groups, the critical values would be correspondingly smaller. For the 101-person sample, for example, the critical value to exceed would be only .17. Differences among coefficients would have further criterion differences depending on whether a particular comparison involved matched or unmatched groups. Although it appears that any particular comparison required can be made from the data reported here, it is not readily apparent how all possible comparisons could be presented in tabular form for easy extrapolation and interpretation.

Further study with other tests and larger examinee samples would be of interest to pursue additional related questions. Among these would be whether context-dependedness of certain item formats exhibit consistent constraints on reliability estimates of various kinds, or on the capacity of items prepared according to those formats to satisfy local-independence assumptions. It would also be useful to conduct further empirical inquiry with larger and more representative samples than those included here regarding the accuracy of current equating procedures. This would further ensure that scaled scores as now derived permit similar score-level interpretations across versions and subsections of the test.

References

Brown, H. D. (1972). Cognitive pruning and second language acquisition. Modern Language Journal, 56, 218-222.

Brown, H. D. (1987). Principles of language learning and teaching (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Gulliksen, H. (1987). Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hale, G. A., Angelis, P. J., and Thibodeau, L. A. (1980). Effects of item disclosure on TOEFL performance (TOEFL Research Report Number 8). Princeton, NJ: Educational Testing Service.

Henning, G. (1982). Growth-referenced evaluation of foreign language instructional programs. TESOL Quarterly, 16(4), 467-477.

Henning, G. (1987). A guide to language testing: Development, evaluation, research. Rowley, MA: Newbury House, 1987.

Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to TOEFL vocabulary test items (TOEFL Research Report Number 35). Princeton, NJ: Educational Testing Service.

Henning, G. (1991). TOEFL subtest functioning. An unpublished statistical report prepared for the TOEFL Committee of Examiners. Princeton, NJ: Educational Testing Service.

Magnusson, D. (1966). Test theory. Reading, MA: Addison-Wesley.

Novick, M. R. (Chair) & the Committee to Develop Standards for Educational and Psychological Testing. (1985). Standards for educational and psychological testing. Washington, DC: The American Psychological Association, Inc.

Sireci, S. G., Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28(3), 237-247.

Stiggins, R. J. & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. Educational Measurement: Issues and Practice, 2(1), 6-10, 25.

Wilson, K. M. (1987). Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language (TOEFL Research Report Number 22). Princeton, NJ: Educational Testing Service.

Cover Printed on Recycled Paper

57906-07589 • Y73M.5 • 275566 • Printed in U.S.A.