DOCUMENT RESUME

ED 283 831 TM 870 308

AUTHOR	Wilson, Kenneth M.
TITLE	Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-87-3; TOEFL-RR-22
PUB DATE	Jan 87
NOTE	88p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Achievement Gains; Correlation; English (Second Language); *Foreign Students; Higher Education; *Language Proficiency; *Language Tests; Listening Comprehension; Predictor Variables; Reading Comprehension; *Scores; *Test Interpretation; Test Wiseness
IDENTIFIERS	Testing Centers; *Test of English as a Foreign Language; *Test Repeaters

ABSTRACT

The Test of English as a Foreign Language (TOEFL) measures the English language proficiency of foreign students who apply for higher education in the United States. It is administered according to a strict schedule through the International and Special Center (I&SC) testing program or at ad hoc intervals by institutional users as part of the TOEFL Institutional Testing Program (INST). This study investigated the patterns of test taking and score change for examinees who repeated the TOEFL in an International or I&SC administration within 24 to 60 months after initial testing. Data for examinees who repeated TOEFL within 1 to 12 months following initial testing by institutions participating in the TOEFL INST /ere also analyzed. In both the I&SC and INST samples, TOEFL repeaters registered substantial average net gains in performance. Findings suggested that there may be differences among national-linguistic groups in typical rate of acquisition of proficiency in English. The magnitude of the average score changes found to be associated with test repetition for repeaters, combined with relatively strong differences among national-linguistic groups in both incidence of test repetition and mean score change, indicates that for purposes of summary reporting on TOEFL performance (e.g., in defining basic reference groups), it is important to specify the test-repetition status of the gorups of examinees involved. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED283831

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 22
JANUARY 198

Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language

Kenneth M. Wilson

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H.C. Wiedenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ERIC
Full Text Provided by ERIC

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 this 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 (1986-87) members of the TOEFL Research Committee include the following:

Paul J. Angelis (chair)	Southern Illinois University
Kathleen M. Bailey	Monterey Institute of International Studies
Ellen Bialystok	York University
John Haskell	Northeastern Illinois University
Grant Henning	University of California at Los Angeles
Shirley L. Menaker	University of Oregon

Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language

Kenneth M. Wilson

Educational Testing Service
Princeton, New Jersey

RR-87-3

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

Copyright © 1987 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

TOEFL, the TOEFL logo, ETS, and the ETS logo are trademarks of Educational Testing Service, registered in the USA, and in many other countries.

Abstract

This study was principally concerned with the analysis of patterns of test taking and score change for examinees who repeated the Test of English as a Foreign Language (TOEFL) in an International or Special Center Testing (I&SC) administration within 24 to 60 months after initial testing. Data for examinees who repeated TOEFL within 1 to 12 months following initial testing by institutions participating in the TOEFL Institutional Testing Program (INST) were also analyzed. In both the I&SC and INST samples, TOEFL repeaters registered substantial average net gains in performance. Findings suggested that there may be differences among national-linguistic groups in typical rate of acquisition of proficiency in English. The magnitude of the average score changes found to be associated with test repetition for repeaters, combined with relatively strong differences among national-linguistic groups in both incidence of test repetition and mean score change, indicates that for purposes of summary reporting on TOEFL performance (e.g., in defining basic reference groups), it is important to specify the test-repetition status of the groups of examinees involved.

Acknowledgements

This study was made possible by the support of the TOEFL Research Committee. Richard Harrison performed the complex task of collating TOEFL records for examinees tested in International or Special Center administrationsover a five-year period, and he also programmed basic analyses of data for these examinees. Steven Vail carried out similar tasks connected with the analysis of data for examinees tested as part of the TOEFL Institutional Testing Program. Brent Bridgeman, Spencer Swinton, and Charles Stansfield read and commented on various versions of the manuscript. These contributions are acknowledged with appreciation. However, the writer assumes responsibility for the contents of this report.

Table of Contents

ERIC Full Text Provided by ERIC

List of Tables and Figures

Item Page

Item Page

SUMMARY

Patterns of Test Taking and Score Change for Examinees Who Repeat The Test of English as a Foreign Language

Kenneth M. Wilson
Educational Testing Service

Each year, thousands of nonnative speakers of English planning undergraduate or graduate study in the United States take the Test of English as a Foreign Language (TOEFL) in order to demonstrate their level of developed proficiency in English. TOEFL provides separate measures of listening comprehension (LC), structure and written expression (SWE), and vocabulary and reading comprehen (V&RC).

TOEFL is offered in scheduled, strictly controlled monthly administrations through the International and Special Center (I&SC) testing program. The test is also administered at ad hoc intervals by institutional users as part of the TOEFL Institutional Testing Program (INST), under conditions controlled by individual institutions. The great majority of examinees are tested in I&SC administrations.

Many of these individuals take TOEFL more than once—that is, they are repeater examinees. Based on a study of the characteristics and test performance of I&SC examinees tested during 1977-1979 (Wilson, 1982), it was estimated that perhaps as many as one third of TOEFL examinees repeat one or more times. Incidence of test repetition was found to vary considerably by country of origin, ranging from less than 10 percent to over 50 percent, and national linguistic contingents with higher percentages of repeating examinees tended to be those with lower mean TOEFL scores. However, the study did not examine questions regarding characteristic patterns of test taking and score change for the examinees who repeated.

Some Working Assumptions Regarding Test Repetition and Score Change for TOEFL I&SC Repeaters

It is assumed that first-time I&SC test takers whose English proficiency as reflected in scores on TOEFL falls below some personal or external criterion level will continue to take TOEFL periodically until they either meet the criterion level or, failing to do so, drop out of the TOEFL population. It is also assumed (a) that between testings they will attempt to improve their proficiency in English through formal and/or informal study/practice and (b) that, on the average, some improvement in English proficiency is likely to occur among repeating examinees between test administrations—improvement that will be reflected in average increases in TOEFL scores. Increased "test-wiseness" and familiarity with the specific test format, statistical factors such as regression to the mean, and so on, are factors that should be kept in mind in evaluating observed changes.

Study Objectives

The study was designed primarily to obtain answers to questions such as the following regarding test taking behavior and "long-term" score change for cohorts of first-time examinees in International and Special Center (I&SC) test administrations:

(a) What are the test-taking patterns of cohorts of first-time I&SC test-takers from different native country and language groups? What is the incidence of 1-time, 2-time, . . ., t-Time test taking?

(b) What are the patterns of change in mean TOEFL performance for repeating examinees by number of times tested and by country of origin?

(c) What is the relationship between change in TOEFL score and variables such as time interval between the initial and last test administrations (within the study period), number of times tested, age, sex, test center (U.S. vs other), educational level (undergraduate vs graduate), and so on?

(d) Which of these variables contribute most to prediction of last-time score after taking first-time score into account?

Study Data

The study focused primarily on data for examinees tested for the first time in a scheduled, monthly TOEFL International and/or Special Center test administration between July 1977 and June 1980, who had accumulated at least one additional test record in an I&SC test administration as of June 1982. The follow-up period—ranging from a minimum of two years to a maximum of five years following initial testing—was deemed to be long enough to cover the tenure of most first-time test takers as "potential participants in an I&SC test administration."

Analysis of long-term test-taking behavior and score change for I&SC repeaters was supplemented by analysis of data from TOEFL Institutional Testing program files for a sample of approximately 10,000 individuals who were identified as first-time institutionally-tested TOEFL examinees, with testing dates between July 1984 and August 1985. For these INST examinees, the maximum time interval for test repetition was about 12 months. The supplementary analysis provided information regarding short-term change in TOEFL performance for individuals presumed to have been enrolled in programs of instruction in English as a second language (ESL) between test administrations.

Procedure

Thirteen analysis groups, based primarily on national origin and associdated linguistic background, were defined for the study. These groups represented a relatively wide range of average performance on TOEFL, incidence of self-reported test repetition, and/or language groups:

Groups known to be characterized by higher incidence of repetition

01 Taiwan
02 Hong Kong
03 Korea
04 Thailand
05 Japan

Groups known to be characterized by medium incidence of repetition

07 Saudi Arabia, Kuwait, Lebanon, Iraq, Jordan, Syria
08 Iran
09 Chile, Colombia, Mexico, Peru, Venezuela
12 Greece, Turkey

Groups known to be characterized by lower incidence of repetition

06 India
10 Germany, Netherlands, Denmark, Norway, Sweden
11 France, Italy, Spain, Portugal
13 Ghana, Nigeria

In the case of group 12 (Greece, Turkey), grouping was based on strong similarities in typical level of performance on TOEFL, incidence of self-reported test repetition, and examinee characteristics other than language (Wilson, 1982). These same groups were employed in analyzing the data on institutionally tested examinees.

For each I&SC analysis group, the study examined (a) the test-taking patterns of first-time examinees (e.g., the number of 1-time, 2-time, . . . , t-Time test takers), (b) the mean change in TOEFL total performance for repeaters generally, and by number of times tested, and (c) mean changes in TOEFL section scores.

For a 20 percent sample of I&SC repeaters, the study examined the relationship between change in TOEFL total score and selected nontest variables: time interval between the initial test administration (t-1) and the last test adminstration (t-T, T= 2, 3, . . ., T), number of times tested, location of the last test center (U.S. vs other), educational level (graduate vs undergraduate), sex, age, and score reporting at time of initial testing (reported TOEFL scores to institutions vs did not do so).

Multiple regression analysis was used to assess the relative contribution of these nontest variables, in combination with initial TOEFL total score, to prediction of the last score of record. The total-sample regression equation was used to compute an expected t-T total score mean for each analysis group. Discrepancies between observed and expected means for the respective analysis groups were computed and evaluated. An evaluation was made of trends across analysis groups in the incidence of test repetition among first-time examinees classified according to score level on TOEFL.

Where feasible and appropriate, parallel procedures were used in analyzing the data for institutionally tested examinees.

Findings Regarding I&SC Repeaters

Test Taking Patterns

Some 28 percent of all first-time examinees between July 1977 and June 1978 had two or more test records in the file as of June 1982. Between 40 and 50 percent of examinees from Taiwan, Hong Kong, Korea, Thailand, and Japan were repeaters; between 19 and 24 percent of those from Iran, Arabic-speaking Mideastern countries, South America and Mexico, Greece, and Turkey repeated; less than 10 percent (between 4 percent and 9 percent) of examinees from India, Ghana and Nigeria, and the European-Germanic and European-Romance groups had more than one test record by the end of the study period.

A majority of I&SC repeaters were one-time repeaters. However, some 12 percent of all test takers were "multiple repeaters"—that is, they took TOEFL three or more times. Incidence of multiple repetition was higher (over 20 percent) for the Asian groups, which had high general incidence of test repeti- tion, and lower for the major European, Indian, and African contingents (less than 1.5 percent). See Tables S-1 and S-2 for data on the the number and percentage of 1-time, 2-time, . . ., T-time test takers, by analysis group.

Repeating examinees, who accounted for only 28 percent of all first-time test takers, accounted for more than one half (about 53 percent) of all test records in the study file for the entire cohort of first-time test takers.

Patterns of Change in Mean Score

In every analysis group, average change in TOEFL total score increased with number of times tested; the amount of change associated with additional testing varied across groups (see Figure S-1).

The first-time TOEFL total means for repeaters classified according to number of times tested during the study period (2, 3, . . ., T-times), varied systematically. Mean time-1 TOEFL total score mean decreased as number of times tested increased. Two-time test takers had higher initial (t-1) means than 3-time test takers, 3-time test takers had higher t-1 means than 4-time test takers, and so on. However, the final (t-T) means of the respective times-tested subgroups were very similar—that is, 2-time, 3-time, 4-time, . . ., t-Time test takers had comparable average time-T total scores. Mean t-T TOEFL total scores typically approached or slightly exceeded 500 in groups with t-1 scores averaging considerably lower than 500. Such regularity suggests that there may be a shared perception of the "acceptable" score level on TOEFL.

Stability of TOEFL Scores

The correlation between the initial and the last observed TOEFL total

Table S-1

Distribution of First-time Test Takers, July 1977 through June 1980, According to Total Number of Test Records Accumulated as of June 1982: By Analysis Group

Group	No. of test-takers	Number of times tested during period										
		1	2	3	4	5	6	7	8	9	10	11+
01	(36950)	20850	8499	3800	1715	895	500	285	182	97	54	73
02	(23278)	13591	4714	2119	1089	631	441	269	190	101	55	75
03	(10178)	5854	2300	1088	489	230	106	62	26	11	5	7
04	(9480)	5640	1943	884	460	280	126	65	41	18	6	17
05	(14308)	7158	3134	1710	910	545	318	187	114	96	43	93
06	(19158)	17442	1425	201	52	17	8	5	3	2	2	1
07	(20839)	16562	2535	992	383	192	88	41	23	11	3	9
08	(31605)	23894	4779	1642	704	352	126	57	22	17	5	7
09	(15021)	12112	1985	607	195	68	25	17	6	2	1	3
10	(6449)	6207	225	15	2	-	-	-	-	-	-	-
11	(6131)	5621	443	54	10	1	2	-	-	-	-	-
12	(7461)	5857	1084	312	109	48	36	9	5	1	-	-
13	(20886)	19621	1125	118	20	1	1	-	-	-	-	-
Total	(221744)	160409	34194	13542	6138	3260	1777	997	612	356	174	285

Table S-2

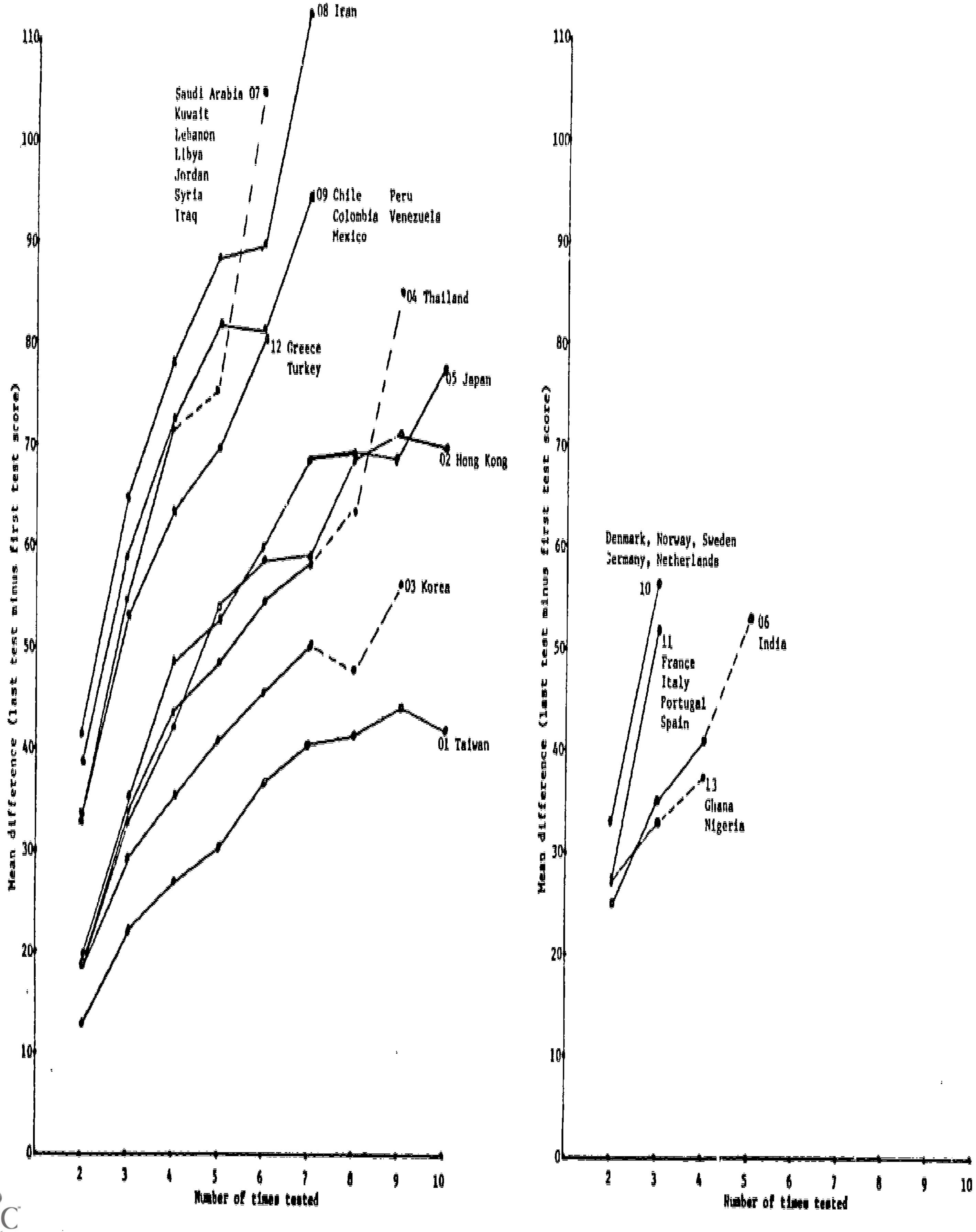
Percentage Distribution of First-time Test takers, July 1977 through June 1980, According to Total Number of Test Records Accumulated as of June 1982: By Analysis Group

Group	No. test-takers	Percent tested designated number of times										
		1	2	3	4	5	6	7	8	9	10	11+
01	(36950)	56.4	23.0	10.3	4.6	2.4	1.4	0.8	0.5	0.3	0.1	0.2
02	(23278)	58.4	20.3	9.1	4.7	2.7	1.9	1.2	0.8	0.4	0.2	0.3
03	(10178)	57.5	22.6	10.7	4.8	2.3	1.0	0.6	0.3	0.1	*	*
04	(9480)	59.5	20.5	9.3	4.9	3.0	1.3	0.7	0.4	0.2	*	0.2
05	(14308)	50.0	21.9	12.0	6.4	3.8	2.2	1.3	0.8	0.7	0.3	0.6
06	(19158)	91.0	7.4	1.0	0.3	0.1	*	*	*	*	*	*
07	(20839)	79.5	12.2	4.8	1.8	0.9	0.4	0.2	0.1	0.1	*	*
08	(31605)	75.6	15.1	5.2	2.2	1.1	0.4	0.2	0.1	0.1	*	*
09	(15021)	80.6	13.2	4.0	1.3	0.5	0.2	0.1	*	*	*	*
10	(6449)	96.2	3.5	0.2	*	-	-	-	-	-	-	-
11	(6131)	91.7	7.2	0.9	0.2	0.2	*	*	-	-	-	-
12	(7461)	78.5	14.5	4.2	1.5	0.6	0.5	0.1	*	*	-	-
13	(20886)	93.9	5.4	0.6	0.1	*	*	-	-	-	-	-
Total	(221744)	72.3	15.4	6.1	2.8	1.5	0.8	0.4	0.3	0.2	0.1	0.1

Note: Analysis groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudi Arabia, Kuwait, Lebanon, Libya, Jordan, Syria, Iraq), 08 (Iran), 09 (Venezuela, Mexico, Colombia, Peru, Chile), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Spain, Italy, Portugal), 12 (Greece, Turkey), 13 (Ghana, Nigeria).

* Less than 0.5 percent.

Figure S-1. Plot of differences between t-1 and t-T TOEFL total score means for repeaters by number of times tested: By analysis group



score, over t-1 to t-T time intervals averaging about one year, was .72 based on data for a 20 percent sample of repeaters without regard to analysis group. This may be compared with internal consistency reliability coefficients of .9 that are routinely obtained for this score and a test-retest coefficient of .9 that has been reported for a sample of examinees retested with an alternate test form within one week. Coefficients reflecting the relative standing of examinees, t-1 to t-T, on TOEFL sections were .56, .63, and .70, for Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary, respectively. Listening comprehension was the least stable and reading comprehension and vocabulary the most stable of the component skills measured by the TOEFL.

These findings, as well as those regarding patterns of change in mean scores, are consistent with the assumption of growth in the skills being measured by the TOEFL for I&SC repeaters between initial and final testing.

Selected Correlates of Total Score Change

In analyses based on a 20 percent sample of repeaters, the same set of four nontest variables was found to contribute most to prediction of change in TOEFL total score between t-1 and t-T in analyses with and without control for t-1 score. These were **number of times tested (Numtsts), time interval in months between t-1 and t-T (Interval), location of test center in the U.S. vs other location (Center-U.S.), and age.** With some variation in detail, the pattern of findings was similar across I&SC analysis groups.

Based on analyses in the total 20 percent sample, weights reflecting the relative contribution (in standard units) of the seven nontest variables to prediction of total net change, when t-1 score was controlled, were as follows:

Variable	Standardized regression weight
Interval in months	.152
Center: U.S.,1; other,0)	.142
Number of times tested	.115
Age at t-1	-.119
Sex (male,1; female,0)	-.065
Educational level (graduate,1; undergraduate,0)	.016
Reported scores at t-1 (yes,1; no,0)	-.008

These findings are consistent with the assumption of growth in English proficiency for repeaters--for example, more time and "effort" (which "times tested" may be thought of as reflecting, at least in part) led to greater change. Other things being equal, younger repeaters and those tested in the U.S. tended to gain more, on the average, than did their older counterparts tested elsewhere.

The weighting of the Center-U.S. variable is consistent with the logical proposition that being in a native-English speaking environment is conducive to improvement in English proficiency.

Predicting t-T Total Score Using t-1 Score and Nontest Variables

In the total 20 percent sample of repeaters (N = 11,612), an equation was developed (with raw score weights) for predicting t-T score, using t-1 total score and scores on the several nontest variables as predictors:

$$\begin{aligned} \text{t-T (predicted)} = {} & .737\ (\text{t-1}) + 3.416\ (\text{Numtsts}) + .681\ (\text{Interval}) \\ & + 12.637\ \text{Center-U.S.} + 1.369\ (\text{Edlevel}) - 1.107\ (\text{Age}) \\ & - 5.877\ (\text{Sex}) - .647\ (\text{Reports}) + 167.391. \end{aligned}$$

This equation was applied to the predictor means for the 13 analysis groups to obtain predicted t-T means. Table S-3 shows the observed and predicted t-1 and t-T means, and the discrepancy between the two, by group. Observed t-T means were 10 or more points higher than predicted for four analysis groups: namely, the European-Romance, South American and Mexican, Greek and Turkish, and European-Germanic repeaters. The average repeater in these four groups had t-T scores that were 10 or more points better than expected, taking into account the t-1 score and scores on all the nontest variables under consideration. Examinees from Taiwan and Thailand had observed t-T means that were some 4 to 7 points lower than expected.

These results may be due to differences among the analysis groups in the amount, intensity, and/or quality of the total English-language intervening experience for the repeating examinees. Such findings may also reflect, to some extent, differences in the degree of linguistic distance between English and the various languages involved, and associated differences in usual rates of acquisition of proficiency in English as a second language.

Tendency to Repeat, by Initial Score Level

It was assumed at the outset that the tendency to repeat TOEFL varies inversely with initial score level. Findings bearing most directly on this assumption are summarized in Figure S-2, which plots the percentage of repeaters among first-time test takers classified in 20-point t-1 TOEFL total-score intervals, for several consolidated analysis groups.

From Figure S-2, it is evident that the tendency to repeat (a) was not a simple linear inverse function of initial score level in any analysis group and (b) was strongly associated with analysis group membership, per se. Within analysis groups, the inverse relationship tended to hold only for first-time test takers with higher initial scores; the percentage of repeaters actually tended to decrease with score level for first-time test takers with lower initial scores.

Figure S-2 indicates clearly that the tendency to repeat was strongly related to analysis group membership. For example, at a t-1 score level of

Table S-3

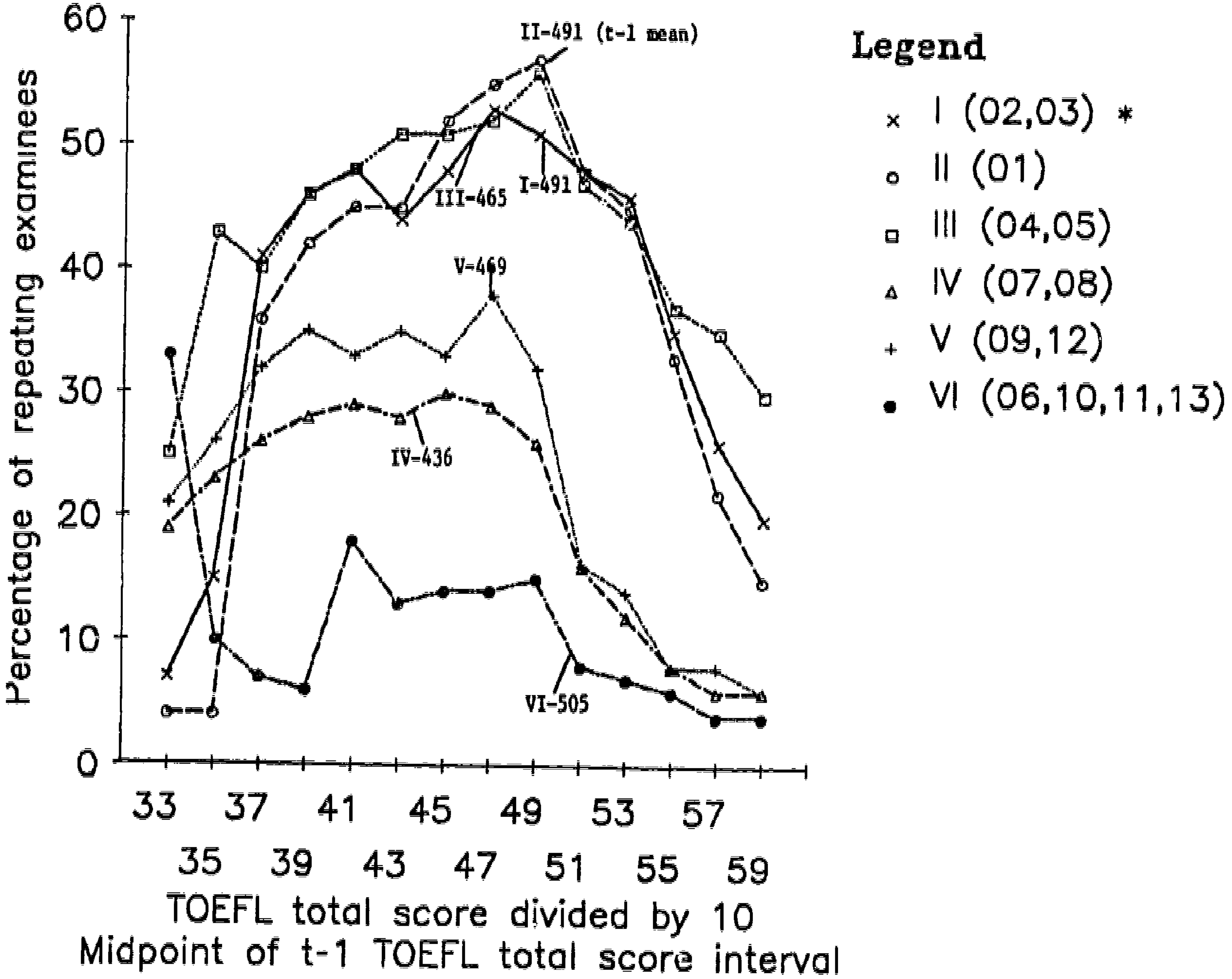
Discrepancies Between Observed and Expected Time-T TOEFL Total Scores, by Analysis Group, When Expectation is Based on an All-Repeater Regression Equation Employing Time-1 Score and Seven Nontest Variables*

Group	N	(1) Time 1 actual (mean)	(2) Time-T actual (mean)	(3) Time-T expected (mean)	Difference (2) - (3)
11	101	530.7	566.9	542.9	+ 24.0
09	566	465.8	513.9	501.5	+ 12.4
12	307	475.0	518.0	507.6	+ 10.4
10	48	531.7	555.7	544.1	+ 11.6
08	1500	434.4	488.6	483.0	+ 5.6
03	819	495.5	521.1	518.8	+ 2.3
06	313	501.2	526.0	526.0	+ 0.0
13	214	490.9	517.9	518.7	- 0.8
05	1346	470.8	503.1	504.2	- 1.1
02	1781	489.5	519.0	520.1	- 1.1
07	814	439.7	485.4	487.3	- 1.9
01	3070	490.7	508.8	512.7	- 3.9
04	733	454.4	485.1	492.5	- 7.4
Total	**11612**	**474.6**	**506.8**	**506.8**	**0.0**

Note: Analyses are based on a 20 percent sample of all repeaters. Groups, ordered in terms of actual minus expected means, are 11 (France, Italy, Portugal, Spain), 09 (Chile Colombia, Mexico, Peru. Venezuela), 12 (Greece, Turkey), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 08 (Iran), 03 (Korea), 06 (India), 13 (Ghana, Nigeria), 05 (Japan), 02 (Hong Kong), 07 (Saudi Arabia, Kuwait, Lebanon, Iraq, Jordan, Syria), 01 Taiwan, 04 (Thailand).

* Expected t-T = .737 (t-1) + 3.416 (Numtsts) + .681 (Intrvl)
+ 12.637 (Center) + 1.369 (Edlevl) - 1.107 (Age)
- 5.877 (Sex) - .647 (Reports) + 167.391.

Figure S-2. Percentage of repeaters by initial score level on TOEFL for consolidated analysis groups (20 percent sampie)



*Entries in ( ) are original analysis groups (see Figure S-1).

500-519, the percentages of repeaters among first-time test takers in the far-Eastern contingents were almost six times greater than the percentage repeating among first-time test takers in the Indian, African, and major European contingents ("Europe" includes analysis groups 10 and 11--see page S-3).

Evaluation of Findings Regarding I&SC Repeaters

The I&SC repeater findings indicate substantial variation among groups of examinees differing in national origin and linguistic background with respect to percentage of repeaters, incidence of multiple repeaters (examinees tested three or more times), and average net gain between testings.

Generally speaking, it is reasonable to infer from the study findings (a) that, when administered to nonnative speakers, TOEFL is measuring developing English-proficiency-related abilities and skills, and (b) that, on the average, the examinees who repeat TOEIL in International and Special Center test administrations experience real improvement in those skills over time-1 to time-T intervals averaging approximately one year.

The study did not control for specific intervening-experience variables such as the nature, amount, duration, intensity, and/or overall "quality" of exposure to the English language. Accordingly, the observed mean changes in TOEFL score should be thought of as indicating net gains without regard to the nature of intervening experience--gains that may be expected to occur for similar groups of I&SC examinees under the range of conditions (a) that prevail between retestings for repeaters in International and Special Center test administrations and (b) that ordinarily influence the tendency of examinees to repeat the test in an I&SC adminstration.

Time interval between initial and final testings and number of times tested (reflecting "effort," at least in part, plus increased "test-wiseness," financial resources, and so on) were correlated positively with change, as was being in the United States at time of final testing (a variable thought of as reflecting degree of immersion in English); age was negatively correlated with change. The other nontest variables (sex, educational level, and reporting versus nonreporting of time-1 scores to institutions) contributed only slightly to the prediction of change or time-T total score.

Study findings indicate that the tendency to repeat TOEFL is not a simple inverse linear function of initial score level. Rather it appears to be strongly related to factors associated with linguistic-cultural background--for first-time examinees at comparable score-levels on TOEFL, the tendency to repeat varied markedly across analysis groups. Why this should be so is not clear. Further study would appear to be warranted.

During the period covered by this study, a minority of first-time I&SC examinees who repeated TOEFL in I&SC administrations generated a majority of all test records: repeaters generated more than half of all test records accumulated by the cohorts of test takers included in the study file, although they represented only 28 percent of all first-time test takers in these cohorts. The concentration of multiple test records varies by country of origin. The magnitude of the average score changes associated with test repetition and the presence of relatively large differences among national-linguistic groups in mean score change suggest that, for purposes of summary reporting of the TOEFL performance of various groups, it is important to specify the test-repetition status of the examinees in various reference groups.

Supplementary Findings: Short-Term Score Change for Repeaters Tested by Institutions

By inference, most institutions that participate in the TOEFL Institutional Testing program use TOEFL for placing students in programs of ESL instruction and/or for evaluating their progress after instruction. Accordingly, it is reasonable to assume that most examinees who repeat the TOEFL under institutional auspices are engaged in formal ESL programs and that observed average changes in TOEFL for samples of INST repeaters may be thought of as average gains associated with the typical range of instruction in English over given periods of time.

To cbtain infcrmation regarding short-term score change for repeating examinees in the Institutional Testing program, supplementary analyses were undertaken. Some 10,000 individuals were identified as having taken TOEFL for the first time in an institutional test administration between July 1984 and August 1985. About 10 percent of these individuals accumulated at least one additional test record during that period; of the repeaters, about 80 percent had only one additional record in the study file for the period under consideration. The mean number of tests per repeater was 2.4, and the average time interval between the initial (time 1 or t-1) test administration and the last observed record (time T or t-T) was 4.2 months. Some 85 percent of the repeaters were tested by institutions located in the United States.

Both first-time test takers generally and those who repeated during the study period had lower TOEFL scores than did the I&SC examinees studied, consistent with the assumptions regarding the enrollment of institutional examinees in ESL instruction; I&SC examinees include individuals at all levels of English proficiency.

For 954 INST repeaters with data on all study variables, the average amount of change in TOEFL Total score, over t-1 to t-T intervals averaging slightly more than four months, was 36 scaled score points. The nontest variables contributing most to prediction of t-T TOEFL total score (and by inference to change, t-1 to t-T) for the INST repeaters were time interval between t-1 and t-T and being enrolled in a U.S. rather than a non-U.S. institution when tested.

As was true for the I&SC repeaters, mean change in TOEFL total was greater for INST repeaters in some national-linguistic analysis groups than in others. Wheuner based on mean change in TOEFL total score or the mean difference between observed and expected t-T scores, the ranks of national-linguistic groups among INST repeaters were found to correspond closely to the ranks for the same national-linguistic groups among examinees identified as I&SC repeaters.

These consistencies in patterns of differences among analysis groups were found even though most of the INST analysis groups were quite small and the circumstances associated with test repetition were not comparable for institutionally tested and I&SC-tested repeaters. Such consistencies strengthen an inference that there are differences among national-linguistic groups in characteristic rate of acquisition of proficiency in English as a second language, at least as measured by the TOEFL.

To the extent that the institutional repeaters were taking special instruction in English as a second language as assumed, the pattern of differences in average gains in TOEFL performance for INST analysis groups in these samples may be thought of as indicative of the pattern of differences that might be expected for similar groups of examinees who may participate in "typical" programs of special ESL instruction, under "typical" program conditions, over time intervals averaging about four months. Of course, the fact that most of the analysis groups were relatively small limits the accuracy of inferences from these data regarding specific average gains to be expected for various groups.

Research is needed to obtain more detailed information regarding the actual circumstances of individuals who repeat the TOEFL in institutional test administrations, the characteristics of instructional programs involved, the conditions of test administration, and so on. Assuming the availability of such information, research involving test-retest data from TOEFL institutional program files could be conducted with ESL programs as units of analysis, with control for, say, duration of ESL program, patterns of instruction, emphasis on English for general purposes as opposed to English for specific academic or occupational purposes, and so on. A capability to track individuals between I&SC and INST test administrations, if developed, would make possible greater control over the previous experience of individuals in studies concerned with growth in English proficiency among institutionally tested examinees.

Patterns of Test Taking and Score Change for Examinees Who Repeat The Test OF English as a Foreign Language

Kenneth M. Wilson
Educational Testing Service

Section I. Introduction

Each year, thousands of non-native speakers of English (largely international students planning undergraduate or graduate study in the United States) take the **Test of English as a Foreign Language** (TOEFL) in order to demonstrate selected aspects of their acquired proficiency in English. TOEFL provides separate measures of Listening Comprehension **(LC)**, Structure and Written Expression **(SWE)**, and Reading Comprehension and Vocabulary **(RC&V)**. A total score is also reported.

Many of these individuals take TOEFL more than one time--that is, they are repeater examinees. Based on a study of the characteristics of TOEFL examinees during 1977-79 (Wilson 1982), it is estimated that about one-third of all examinees take TOEFL more than one time, either in regular International and Special Center test administrations or in institutionally contolled test administrations. The incidence of test repetition varied considerably by country of origin, ranging from under 10 percent to over 50 percent. National contingents with higher percentages of repeaters tended to be those with lower mean TOEFL scores--a correlation of -.64 was found between these two summary statistics in data for 129 native-country groups.

Findings such as these point up the importance of the repeated test-taking or repeater phenomenon. However, there have been no systematic analyses either of the characteristics and test taking behavior of the repeating examinees or of changes in TOEFL performance across repetitions.

Some Working Assumptions

It is assumed that the tendency to repeat TOEFL is inversely associated with the initial score level attained by an examinee. First-time test takers in International and Special Center (I&SC) test administrations, without regard to country of origin, whose English proficiency as reflected in TOEFL scores falls below some personal or external criterion level are likely to continue to take TOEFL periodically until they either meet the criterion level or, failing to do so, drop out of the TOEFL I&SC examinee population.

It is further assumed (a) that between testings many, if not most, TOEFL repeaters attempt to improve their proficiency in English in a variety of ways--through independent study, tutoring, intensive study of English as a second language in a formal program, and so on, and (b) that, on the average, some improvement in English proficiency is likely to occur among repeating test takers between successive test administrations--improvement that will be reflected in average increases in TOEFL scores for groups of repeaters.

Some observed improvement in test performance might be due to factors other than improved proficiency in the English language skills measured by TOEFL--for example, increased familiarity with the specific test format, generally increased "test-wiseness," statistical factors such as regression to the mean, and so on. Factors such as these should be kept in mind in evaluating observed changes in TOEFL scores for repeating I&SC examinees.

Study Objectives

This study was designed (a) to identify and evaluate the major patterns of TOEFL-taking for cohorts of first-time examinees in International and Special Center (I&SC) test administrations, (b) to document average change in performance on TOEFL for the repeating candidates, and to study the relationship of variables such as number of repetitions, time interval between testings, and initial score level, to score change, (c) to study the correlations between initial and subsequent testings, and (d) to assess the contribution of nontest variables in combination with initial score to prediction of final observed standing on the TOEFL.

The average score changes observed for TOEFL repeaters under the range of conditions that ordinarily prevail between test administrations for I&SC examinees who repeat TOEFL may be thought of as providing base line perspective on gain over time without regard to intervening variables for individuals who are selected into the TOEFL I&SC repeater population.

In more specific terms, the study was designed to obtain answers to questions such as the following:

(1) What are the test-taking patterns of cohorts of first-time I&SC test-takers from different native country and language groups? What is the incidence of one-time, two-time, . . . , t-time test taking?

(2) What are the characteristic patterns of change in mean TOEFL performance for two-time, three-time, . . ., t-time test takers from different native country and language groups?

(3) What is the relationship between change in TOEFL score and such variables as time interval between initial and last test administrations, number of times tested, age, sex, test center (U.S. vs other), educational level (undergraduate vs graduate), and so on?

(4) Which of these variables contribute most to prediction of last-time scores after taking first-time TOEFL scores into account?

These questions pertain primarily to examinees tested in I&SC test administrations. The study was also concerned with obtaining evidence regarding short-term changes in TOEFL scores among repeaters tested by institutions.

Section II. Study Data and Procedures

Data for the study were obtained from test files for 60 International and Special Center test administrations between July 1977 and June 1982, inclusive. These files were consolidated to form a "history file"--that is, multiple test records for individuals were collated by using name, sex, and date of birth matching criteria. For individuals with more than one test record according to the matching criteria, records were ordered in terms of date of test administration, from the earliest to the most recent.

Following the organization of records to reflect the history of test taking for each individual in the file, first-time test takers during the three years between July 1977 and June 1980 were identified using responses to a standard question on previous experience with TOEFL. Individuals tested during one of the first 36 test administrations were classified as first-time test takers if in registering for the initial test of record during that period they reported no previous experience with TOEFL. Only the records of first-time test takers who planned undergraduate or graduate study in the U.S. or Canada (the primary TOEFL populations) were included in the study.

The study file included all the test records of first-time test takers during the period July 1977 through June 1980 that had accrued as of June 1982. Thus, the maximum time interval for a potential repeater varied between two years, for first-time test takers in June 1980, and five years, for those tested initially in July 1977.

Forming Analysis Groups

Given the heterogeneity of the TOEFL examinee population with respect to national origin and linguistic background, and the known differences by country in incidence of self-reported test repetition (Wilson 1982), it was decided to select for analysis data for examinees from several countries or groups of countries representing a range of average performance on TOEFL and incidence of self-reported test repetition, and/or major language groups. The **analysis groups** defined for the study were as follows:

01 Taiwan
02 Hong Kong
03 Korea
04 Thailand
05 Japan
06 India
07 Saudi Arabia, Kuwait, Lebanon, Iraq, Jordan, Syria
08 Iran
09 Chile, Colombia, Mexico, Peru, Venezuela
10 Germany, Netherlands, Denmark, Norway, Sweden
11 France, Italy, Spain, Portugal
12 Greece, Turkey
13 Ghana, Nigeria.

Examinees from Greece and Turkey were grouped for analysis because of strong similarities in level of performance on TOEFL, known incidence of test repetitition, and other examinee characteristics (see Wilson, 1982, Tables 2.1, 3.1, 4.1, and 5.1). Students from these two countries also tend to exhibit similar patterns of performance on the Graduate Record Examinations General Test (see Wilson, 1984, Table 13 and Figure 6.1). With the exception of examinee contingents from India and Ghana-Nigeria (analysis groups 06 and 13, respectively), which are heterogeneous with respect to native language, examinees from the selected native countries tend to be relatively homogeneous linguistically (Wilson, 1982).

Supplementary Data on Institutionally Tested Repeaters

It is assumed that institutions participating in the TOEFL Institutional Testing program (INST) use TOEFL to identify individuals needing ESL instruction and/or to assess their progress after a period of ESL instruction. Thus, changes between initial and subsequent testings for individuals identified as institutionally tested repeaters may be thought of as reflecting changes associated with the typical range of ESL instruction offered by the institutions involved over given time periods.

Using the same criteria for first-time test takers that were used with the I&SC data, some 10,000 individuals were identified as having taken TOEFL for the first time in an institutional test administration between July 1984 and August 1985. About 10 percent of these first-time institutional test takers repeated TOEFL at least once during the study period. Procedures employed in analyzing data for institutionally tested repeaters and related findings are described in detail in Section X.

Section III. Test-Taking Patterns of First-Time I&SC Examinees

The 13 analysis groups defined for the study included a total of 221,784 individual examinees who took TOEFL for the first time in an I&SC administration between July 1977 and June 1980, inclusive. The distribution of these examinees according to repeater versus nonrepeater status and analysis group is shown in Table 1. As of June 1982, 61,355 examinees (27.7 percent) had repeated TOEFL one or more times and 160,409 (72.3 percent) had not done so (were classified as one-time test takers, or nonrepeaters).

o Among the 13 groups, incidence of repeated test taking varied from 3.8 percent to 50 percent.

o Five Asian contingents (Taiwan, Hong Kong, Korea, Thailand, and Japan) registered the highest incidence of repeated test taking: between 40 and 50 percent had repeated at least once by June 1982. Moderate incidence of repeating (between 19 and 24 percent) was found for four contingents (the Arabic- and Farsi-speakers from the Mideast, Spanish-speakers from the Americas, and the Greek-Turkish contingent). Low incidence of repeated test taking (less than 10 percent) was found for groups 06, 10, 11, and 13 (the Indian, the major European, and the principal African contingents).

More detailed data on test taking patterns for these analysis groups are provided in Table 2 and Table 3 which show, respectively, the number and the percentage of first-time examinees between July 1977 and June 1980 who had accumulated designated numbers of test records as of June 1982. From the data in Table 2 it may be determined that the 61,355 repeaters, who accounted for only about 28 percent of all first-time test takers during the study period, accounted for a total of over 180,000 test records, or about 53 percent of the total of over 340,000 test records. Only 160,000+, or about 47 percent of the test records for the period were accounted for by one-time test takers.

Some examinees (0.1 percent) took TOEFL more than 10 times during the study period but only 12.3 percent of all test takers had more than two records on file (had repeated more than one time) by the end of the study period—that is, were multiple repeaters.

As might be expected, incidence of multiple repetition was greatest for the national-linguistic analysis groups with the highest general incidence of repeaters, and lowest for those with relatively few repeaters: more than 20 percent of the examinees in groups **01, 02, 03, 04,** and 05 were multiple repeaters as compared to fewer than 1.5 percent of those in groups **06, 10, 11,** and **13.** In most of the analyses of test performance in the study, data for the small numbers of 11, 12, . . ., 20+-time test takers were not included since Ns were to small to warrant analysis for these times-tested subgroups.

Table 1

Distribution of First-Time Test Takers During the Period July 1977 through June 1980 According to Repeater versus Nonrepeater Status as of June 1982

Group**	N	Status as of June 1982*			
		Repeater		Nonrepeater	
		No.	Percent	No.	Percent
01	36950	16100	43.6	20850	56.4
02	23278	9687	41.6	13591	58.4
03	10178	4324	41.5	5854	58.5
04	9480	3840	40.5	5640	59.5
05	14308	7150	50.0	7158	50.0
06	19158	1716	9.0	19157	91.0
07	20839	2909	20.5	16562	79.5
08	31605	7711	24.4	23894	75.6
09	15021	2909	19.4	12112	80.6
10	6449	242	3.8	6207	96.2
11	6131	510	8.3	5621	91.7
12	7461	1604	21.5	5857	78.5
13	20866	1265	6.2	19621	93.8
Total	221744	61355	27.7	160409	72.3
Median			21.5		79.5

* 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudi Arabia, Kuwait, Lebanon, Libya, Jordan, Syria, Iraq), 08 (Iran), 09 (Mexico, Chile, Colombia, Peru, Venezuela), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Italy, Spain, Portugal), 12 (Greece, Turkey), 13 (Nigeria, Ghana).

** Repeaters are first-time test takers, July 1977 through June 1980, who had at least one additional test record in the file as of June 1982; nonrepeaters had only one test record in the file.

Table 2

Distribution of First-time Test Takers, July 1977 through June 1980, According to Total Number of Test Records Accumulated as of June 1982: By Analysis Group

Group	No. of test-takers	Number of times tested during period										
		1	2	3	4	5	6	7	8	9	10	11+
01	(36950)	20850	8499	3800	1715	895	500	285	182	97	54	73
02	(23278)	13591	4714	2119	1089	631	441	269	190	101	55	75
03	(10178)	5854	2300	1088	489	230	106	62	26	11	5	7
04	(9480)	5640	1943	884	460	280	126	65	41	18	6	17
05	(14308)	7158	3134	1710	910	545	318	187	114	96	43	93
06	(19158)	17442	1425	201	52	17	8	5	3	2	2	1
07	(20839)	16562	2535	992	383	192	88	41	23	11	3	9
08	(31605)	23894	4779	1642	704	352	126	57	22	17	5	7
09	(15021)	12112	1985	607	195	68	25	17	6	2	1	3
10	(6449)	6207	225	15	2	-	-	-	-	-	-	-
11	(6131)	5621	443	54	10	1	2	-	-	-	-	-
12	(7461)	5857	1084	312	109	48	36	9	5	1	-	-
13	(20886)	19621	1125	118	20	1	1	-	-	-	-	-
Total	(221744)	160409	34194	13542	6138	3260	1777	997	612	356	174	285

Table 3

Percentage Distribution of First-time Test Takers, July 1977 through June 1980, According to Total Number of Test Records Accumulated as of June 1982: By Analysis Group

Group	No. test-takers	Percent tested designated number of times										
		1	2	3	4	5	6	7	8	9	10	11+
01	(36950)	56.4	23.0	10.3	4.6	2.4	1.4	0.8	0.5	0.3	0.1	0.2
02	(23278)	58.4	20.3	9.1	4.7	2.7	1.9	1.2	0.8	0.4	0.2	0.3
03	(10178)	57.5	22.6	10.7	4.8	2.3	1.0	0.6	0.3	0.1	*	*
04	(9480)	59.5	20.5	9.3	4.9	3.0	1.3	0.7	0.4	0.2	*	0.2
05	(14308)	50.0	21.9	12.0	6.4	3.8	2.2	1.3	0.8	0.7	0.3	0.6
06	(19158)	91.0	7.4	1.0	0.3	0.1	*	*	*	*	*	*
07	(20839)	79.5	12.2	4.8	1.8	0.9	0.4	0.2	0.1	0.1	*	*
08	(31605)	75.6	15.1	5.2	2.2	1.1	0.4	0.2	0.1	0.1	*	*
09	(15021)	80.6	13.2	4.0	1.3	0.5	0.2	0.1	*	*	*	*
10	(6449)	96.2	3.5	0.2	*	-	-	-	-	-	-	-
11	(6131)	91.7	7.2	0.9	0.2	0.2	*	*	-	-	-	-
12	(7461)	78.5	14.5	4.2	1.5	0.6	0.5	0.1	*	*	-	-
13	(20886)	93.9	5.4	0.6	0.1	*	*	-	-	-	-	-
Total	(221744)	72.3	15.4	6.1	2.8	1.5	0.8	0.4	0.3	0.2	0.1	0.1

Note: Analysis groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudi Arabia, Kuwait, Lebanon, Libya, Jordan, Syria, Iraq), 08 (Iran), 09 (Venezuela, Mexico, Colombia, Peru, Chile), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Spain, Italy, Portugal), 12 (Greece, Turkey), 13 (Ghana, Nigeria).

* Less than 0.5 percent.

Section IV. Patterns of Mean Score-Change, by Analysis Group

Table 4 shows the initial (time-1 or t-1) and the final (time-T or t-T) TOEFL total means for repeaters, without regard to number of repetitions, by analysis group. Average change (the differences between t-T and t-1 means) is also shown. Groups are ordered from high to low in terms of incidence of repeaters. The various analysis groups differed in initial and final level of TOEFL performance, as well as mean change.

- The median change (net gain) in TOEFL total score means for the 13 analysis groups was 31.2; average gains ranged from 19.0 (Taiwan, 01) to 53.6 (08, Arabic-speakers from the Mideast).

- The four analysis groups with the lowest percentages of repeaters were among the six with initial test means above the median for all groups; only two of the six groups with above-median percentages of repeaters were above the initial test score median. This is consistent with the assumption of an inverse relationship between initial test score and the tendency to repeat, but some lower-scoring contingents are not high in repeater rate.

- Five of the six groups with above median average gain were among the six groups having initial test means below the median for all groups.

The relationship between initial test score mean and net gain is highlighted in Table 5, which shows initial and final test means and the corresponding differences for the 13 analysis groups, ordered from low to high in terms of mean score on the initial test (t-1). In Table 5, the raw score data from Table 4 have been subjected to a z-scale transformation with respect to the distribution of most recent scores for all TOEFL examinees tested between July 1977 and June 1979 who planned to study in the U.S. or Canada (Wilson, 1982)—that is, deviations from the reference group mean are expressed in terms of reference group standard deviation units.

Adjusted mean gains in TOEFL total score amounted to one-half of a standard deviation or more for the two Mideastern contingents (groups 07 and 08), the Spanish speakers from the Americas (group 09), the Greek and Turkish contingent (group 12), and the French, Italian, Spanish, and Portugese cluster (group 11). The smallest mean adjusted gain (0.24 z-scaled units) was registered by the Taiwanese contingent (01), but the second Chinese-speaking contingent (02, Hong Kong) showed an adjusted gain of 0.41 z-scale units. Adjusted gains for the remaining contingents ranged from 0.34 for the Korean contingent (03) to 0.43 for the Japanese repeaters (05).

The data in Tables 4 and 5 point up the marked differences among analysis groups with respect to both initial and final test means—1.42 reference-group standard deviations separated the t-1 means for groups 10 and 8 (the highest and lowest scoring groups). With regard to the t-T means, the range was less (1.15 z-scale units) but still substantial. However, while 10 of the 13 groups of repeaters had t-1 means below the general reference group mean, only three had t-T means below that level.

Table 4

Initial and Final TOEFL Total Means for Repeaters Without Regard to Number of Repetitions

Group*	Repeaters**		Admin	Admin	Difference
	No.#	Percent	t-1	t-T	(t-T)-(t-1)
05	7150	50.0	470.1	504.5	34.4
01	16100	43.6	489.9	508.9	19.0
03	4324	42.5	496.9	522.3	25.4
02	9687	41.6	490.6	521.8	31.2
04	3840	40.5	458.0	488.6	30.6
07	4277	25.8	439.7	484.0	44.3
08	7711	24.4	434.8	488.4	53.6
12	1604	21.5	473.4	515.4	42.0
09	2909	19.4	465.9	512.8	46.9
06	1716	9.0	508.0	533.2	25.2
11	510	8.3	523.2	559.2	36.0
13	1265	6.2	494.5	522.4	27.9
10	242	3.8	538.3	567.5	29.2
Median		24.4	489.9	515.4	31.2

* Groups, ordered in terms of incidence of repeaters, are 05 (Japan), 01 (Taiwan), 03 (Korea), 02 (Hong Kong), 04 (Thailand), 07 (Saudi Arabia, Kuwait, Lebanon, Libya, Jordan, Syria, Iraq), 08 (Iran), 12 (Greece, Turkey), 09 (Mexico, Colombia, Chile, Peru, Venezuela), 06 (India), 11 (France, Spain, Italy, Portugal), 13 (Ghana, Nigeria), 10 (Germany, Netherlands, Denmark, Norway, Sweden).

** Only data for 2-, 3-, . . ., 10-time test takers were included in calculating means.

Ns include repeaters tested more than 10 times (that is, 11-, 12-, . . . , T-time test takers, data for whom were not included in calculating the means reported in this table).

Table 5

Initial and Final TOEFL Total Means for Analysis Groups, Expressed as Deviations from the Most Recent Score Mean for a General Sample of 1977-1979 Examinees, and Mean Differences With and Without Adjustment of the Initial Means for Unreliability

Group*	N	Initial mean (t-1)	Final mean (t-T)	Difference (t-1)-(t-T)	
				Observed	Adjusted**
08	7711	-0.96	-0.23	0.87	0.64
07	4277	-0.90	-0.29	0.61	0.52
04	3840	-0.64	-0.22	0.42	0.36
09	2909	-0.54	0.11	0.64	0.59
05	7150	-0.48	0.01	0.47	0.43
12	1604	-0.43	0.14	0.58	0.53
01	16100	-0.21	0.05	0.26	0.24
02	9687	-0.20	0.23	0.43	0.41
13	1265	-0.14	0.24	0.38	0.37
03	4324	-0.11	0.24	0.35	0.34
06	176	0.04	0.39	0.34	0.35
11	510	0.25	0.74	0.49	0.52
10	242	0.46	0.86	0.40	0.41
Median		-0.21	0.14	0.43	0.41

Note. In this table, the scaled score means shown in Table 4 have been expressed as deviations from the most recent score mean of all TOEFL examinees tested between July 1977 and June 1979, in standard deviation units. As reported elsewhere (Wilson, 1982), the mean for this reference group was 505 and the standard deviation was 73. For example, the t-1 total score mean for Group 08 (434.8, from Table 4) was 70.6 scaled-score points, or 0.96 standard deviation units (70.7 / 73.0), lower than the 1977-1979 reference group mean of 505; the t-T mean (488.4) was 16.6 scaled score points (0.23 standard deviation units) below the reference group mean; and so on.

* Groups, ordered from low to high in terms of mean score on the initial test, are: 08 (Saudi Arabia, Kuwait, Jordan, Lebanon, Libya, Iraq, Syria), 07 (Iran), 04 (Thailand), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 05 (Japan), 12 (Greece, Turkey), 01 (Taiwan), 02 (Hong Kong), 13 (Ghana, Nigeria), 06 (India), 11 (France, Italy, Spain, Portugal), 10 (Germany, Netherlands, Denmark, Norway, Sweden).

** In calculating the adjusted differences, the t-1 mean was regressed, using an estimated reliability of .90. Reliability estimates are not available for first-time test takers who repeat. Thus, differences involving the adjusted time-1 means should be thought of primarily as illustrative of "regression effects."

Section V. Mean Score Change by Number of Times Tested

The data in Tables 4 and 5 reflect average differences between the t-1 and t-T TOEFL total scores for repeaters without regard to number of test repetitions. Table 6 shows the t-1 and t-T TOEFL total score means for 2-, 3-, . . ., 10-time test takers in the respective analysis groups. On balance, the overall picture is one of added gains with added testing.

The t-1 means of multiple repeaters varied inversely with number of times tested. Examinees with only two test records had higher t-1 means than those with three, those with three test records had higher t-1 means than those with four, and so on. While there were some minor reversals in this pattern, it was the prevailing one.

Times-tested subgroups did not differ very much with respect to level of t-T performance. For subgroups with t-1 means below 500, t-T means approaching or slightly higher than 500 were typical. In groups 02, 03, 04, and 05 examinees tested five to 10 times had t-T means comparable to those of two-time test takers. In groups 07 and 08, five and seven time test takers had higher t-T means than those tested only two times. For the four groups with very low incidence of repetition (groups 06, 10, 11, and 13), for the high-incidence Taiwanese contingent (group 01), and for groups 09 and 12, there was a modest decline in t-T means across times-tested subgroups.

Table 7 shows differences beteen t-1 and t-T means for examinees tested designated numbers of times. Groups are clustered by overall incidence of test repetition, from high to low.

As expected from the data in Table 6, mean gain tended to increase with additional testing--individuals taking TOEFL three times gained more on the average than those taking TOEFL only twice, four-time test takers gained more than three-time test takers, and so on.

At all levels of repeated test taking, repeaters in analysis groups with high incidence of repetition (01, 02, 03, 04, and 05) showed lower mean gain than repeaters in the four analysis groups with moderate incidence of test repetition (07, 08, 09, and 12). Considering only two-time test takers, the largest repeater subgroup, gains for the high-incidence analysis groups were lowest (ranging between 12.4 and 19.8), those for the moderate-incidence groups were highest (ranging between 31.4 and 41.6), and gains for the low-incidence groups were in-between (ranging from 22.5 to 32.7).

Similarities and differences among the 13 analysis groups in patterns of mean gain for individuals tested 2, 3, . . ., 10-times, as reported in Table 7, are pointed up graphically in Figure 1. Several clusters of analysis groups based on general similarity in gain patterns are discernible: (a) groups 07, 08, 09, and 12, (b) the contingents from Thailand, Hong Kong, and Japan (groups 04, 02, and 05), (c) the Taiwanese and Korean contingents (groups 01 and 03), although these may be thought of as constituting two distinctive patterns rather than a cluster, (d) the Germanic and Romance language (European) contingents (groups 10 and 11), and (e) the contingents

Table 6

First and Last Administration Mean TOEFL Total Scores for Repeaters, by Number of Times Tested and Analysis Group

Group/Admin@		Number of times tested									All
		2	3	4	5	6	7	8	9	10	individ.
01	T-1	502.7	486.0	473.4	465.2	456.0	446.6	450.6	435.8	441.2	489.9
	T-t	515.1	508.0	500.1	495.2	492.5	487.0	492.5	479.6	482.9	508.9
02	T-1	502.3	489.2	478.2	472.1	468.4	466.1	460.5	458.8	451.6	490.6
	T-t	520.6	521.5	520.2	526.0	526.6	524.4	529.6	529.6	521.3	521.8
03	T-1	504.4	496.8	483.9	490.4	468.6	468.4	450.2a	440.8b	455.8a	496.9
	T-t	522.8	525.2	518.8	520.9	513.9	518.2	498.0	496.8	496.0	522.3
04	T-1	465.5	457.7	448.9	444.2	437.89	428.0	434.8a	415.4b	426.5c	458.0
	T-t	485.3	491.4	492.5	493.6	492.3	487.3	498.2	500.2	481.0	488.6
05	T-1	483.3	469.7	460.4	453.3	445.1	440.2	433.8	430.6	432.1a	470.1
	T-t	502.8	504.8	508.6	506.2	504.2	509.7	502.6	499.1	509.0	504.5
06	T-1	515.2	474.4	463.8	480.6b	451.4c	465.4c	*	*	*	508.0
	t-t	537.7	509.4	504.5	533.4	524.2	494.8	*	*	*	533.2
07	T-1	449.3	431.8	421.3	413.3	411.0	421.6a	401.6b	422.6b	*	439.7
	T-t	483.1	484.8	482.8	491.6	491.6	490.6	486.4	520.4	*	484.0
08	T-1	442.1	428.0	420.7	414.0	413.3	397.6	412.1b	394.1b	399.6b	434.8
	T-t	483.7	492.4	498.7	502.3	502.7	509.9	501.4	493.5	499.4	488.4
09	T-1	477.5	447.7	433.9	426.6	415.2a	403.1b	392.3c	*	*	465.9
	T-t	516.0	506.4	506.2	508.4	496.1	497.4	506.2	*	*	512.8
10	T-1	540.1	519.1b	*							538.3
	T-t	567.3	571.0	*							567.5
11	T-1	527.4	499.3	472.6b	*	*	*				523.2
	T-t	560.1	555.8	533.2	*	*	*				559.2
12	T-1	486.4	457.3	437.3	433.5a	414.4a	384.1c	420.6c	*		473.4
	T-t	517.8	511.6	509.0	509.0	509.4	488.6	508.6	*		515.4
13	T-1	496.9	475.2	475.3b	*	*					494.5
	T-t	524.0	508.2	512.8	*	*					522.4

@ T-1 is the mean time-1 test score; T-t is the mean score on the last test of record. The "all individ." entries (last column) are T-1 and T-t means for all individuals who repeated, without regard to number of times. Groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudi Arabia, Kuwait, Lebanon, Jordan, Syria, Iraq), 08 (Iran), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 10 (Germany, Netherlands, Norway, Denmark, Sweden), 11 (France, Italy, Spain, Portugal), 12 (Greece, Turkey), 13 (Ghana, Nigeria).

* Less than 5 cases; a (N = 25-49); b (N = 10-24); c (N = 5-9). See Table 2 for exact Ns, all categories.

Table 7

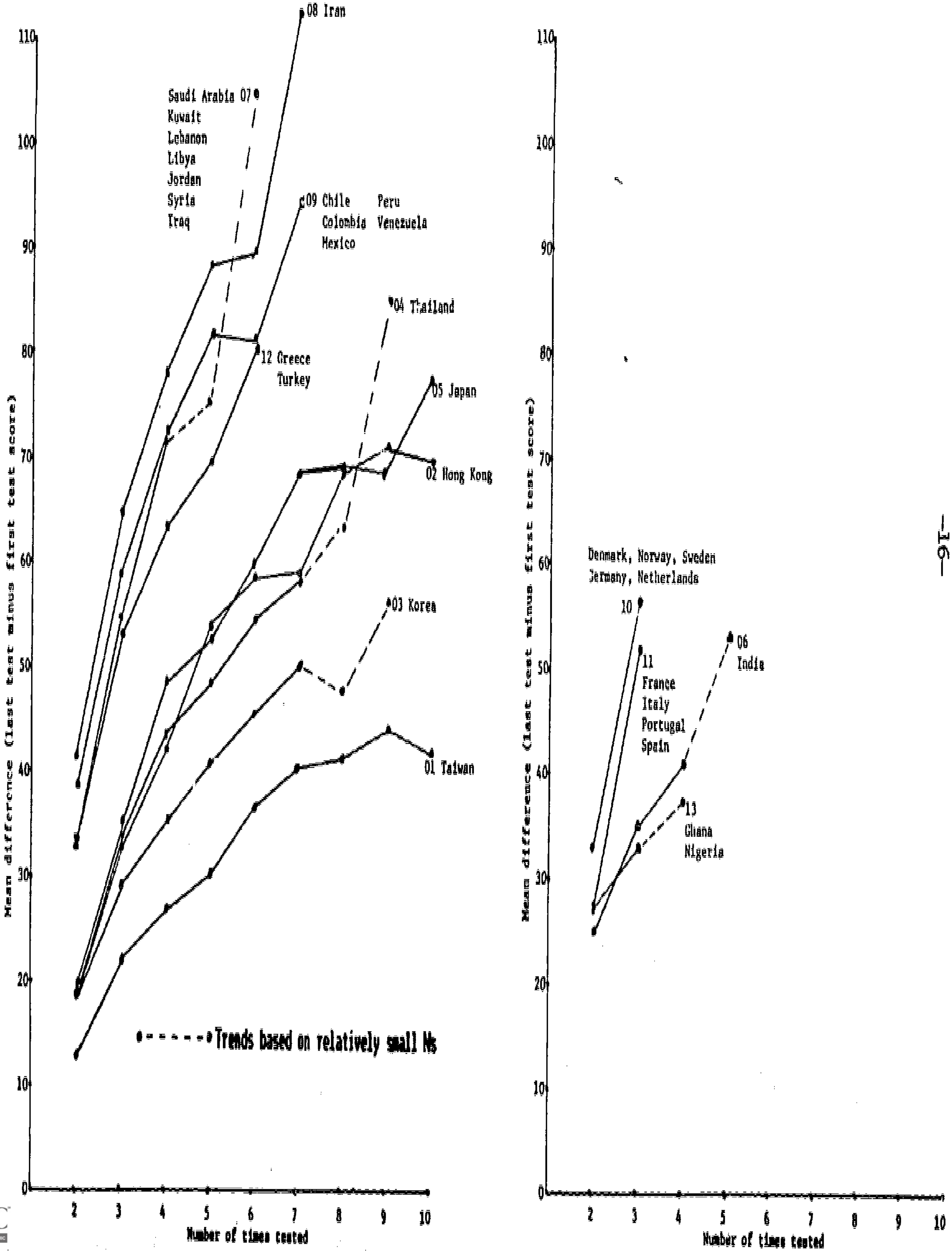
Difference between Initial and Final TOEFL Total Means,
by Number of Times Tested and Analysis Group

Group*	Number of times tested								
	2	3	4	5	6	7	8	9	10
01	12.4	22.0	26.7	30.0	36.5	40.4	41.9	43.8	41.7
02	18.3	32.3	42.0	53.9	58.2	58.3	69.1	70.8	69.7
03	18.4	28.6	34.9	40.5	45.3	49.8	47.8a	56.0b	40.2c
04	19.8	33.7	43.6	49.4	54.5	59.3	63.4a	84.8b	54.5c
05	19.5	35.1	48.2	52.9	59.1	68.5	68.8	68.5	76.9a
07	33.8	53.0	63.3	69.5	80.6	69.0a	84.8b	97.8b	
08	41.6	64.4	78.0	88.3	89.4	112.3	89.3b	99.4b	99.8c
09	38.5	58.7	72.3	81.8	80.9a	94.3b	113.9c		
12	31.4	54.3	71.7	75.5a	95.0a	104.5c	88.0c		
06	22.5	35.0	40.7	52.8b	72.8c				
10	27.2	51.9b							
12	32.7	56.5	60.6b						
13	27.1	33.0	37.5b						

* 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 07 (Saudi Arabia, Jordan, Lebanon, Libya, Iraq, Syria), 08 (Iran), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 12 (Greece, Turkey), 06 (India), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 12 (France, Italy, Spain, Portugal), 13 (Nigeria, Ghana).

a N = 25–49
b N = 10–24
c N = 5–9

Figure 1. Plot of differences between t-1 and t-T TOEFL total means for repeaters in the respective analysis groups, by number of times tested



from India and Ghana-Nigeria (groups 06 and 13).

Contrast in score change for TOEFL and GRE repeaters. A detailed picture of the patterns of t-1 and t-T means for 2, 3, . . . , 10-time test takers in the respective analysis groups is provided in Figure 2. Broken lines connecting t-1 and t-T means are used to indicate breaks in the typical inverse relationship between t-1 mean and number of times tested. For example, in the array for Taiwan, 8-time test takers had slightly higher t-1 means than did 7-time test takers, the t-1 mean for 10-time test takers was somewhat higher than that for 9-time test takers, and so on. The horizontal line for each array indicates the mean TOEFL total score of all first-time test takers (nonrepeaters plus repeaters) in the analysis group.

A plot of t-1 and t-T GRE Verbal Test means for 1, 2, . . ., 5-time GRE test takers, adapted from findings reported by Rock and Werts (1979), is included for comparative purposes (page 1 of figure, lower right frame).

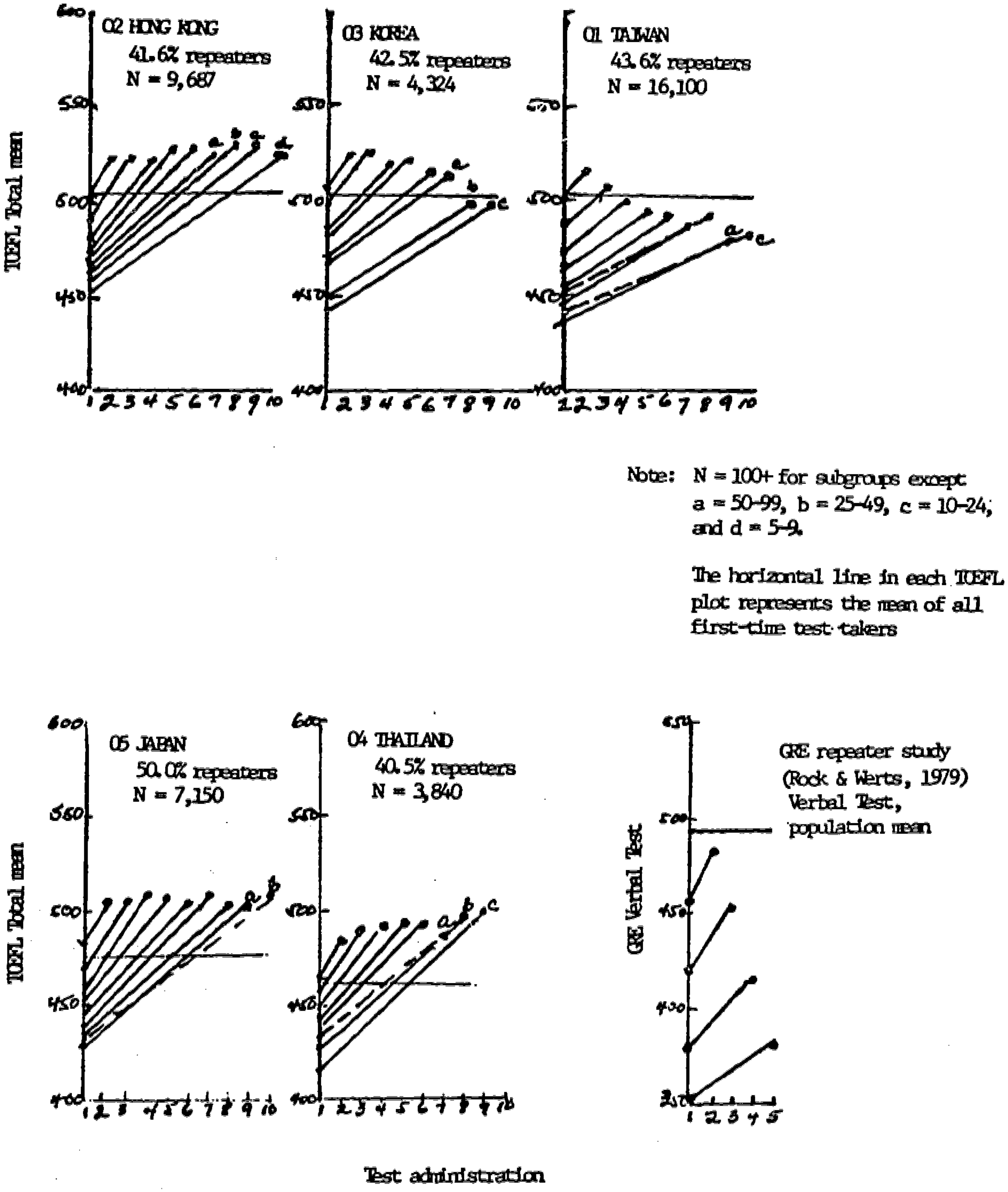
For TOEFL repeaters, a strong inverse relationship between t-1 mean and number of times tested is clearly evident in Figure 2. The general comparability of the t-T means of 2, 3, . . ., T-time test takers is also evident. Among the five high-incidence and the four moderate-incidence contingents, only the time-T means of Taiwanese repeaters show a moderately declining gradient as number-of-times-tested increases. For the contingents with low general incidence of test repetition, and few multiple repeaters, trends are mixed.

Differences in patterns of mean score change on the GRE Verbal Test for GRE repeaters (largely native English speakers) and mean score-change for TOEFL repeaters (nonnative English-speakers) are quite clear in Figure 2. For GRE repeaters, the mean increase in verbal score tended to decline with number of times tested. Thus, the array of t-time means for 2-, 3-, 4-, and 5-time test takers has a sharply declining gradient. The most-recent test mean for 3-time test takers was comparable to the t-1 mean for 2-time test takers, the most recent mean of 4-time test takers is like the initial mean of 3-time test takers, and so on. In addition, none of the GRE times-tested repeater groups had most-recent score means that reached the population value and only the most-recent score mean for 2-time GRE test takers approached the population value.

The patterns of score change for TOEFL multiple repeaters, especially those in the contingents with high and moderate incidence of repetition, contrast sharply with the observed patterns of score change on the GRE Verbal Test for GRE repeaters. For TOEFL repeaters, (a) level of performance on the most recent test administration (t-T) for 2, 3, . . . , t-time test takers appears to be independent of their average levels of t-1 performance, and (b) t-T means tend to equal or surpass the mean for all first-time test takers in the respective groups and most-recent score mean for all TOEFL examinees.

In the low-incidence contingents (groups 06, 09, 10, and 13), some tendency for the t-time means of multiple repeaters to fall below the mean for all first-time test takers is evident in Figure 2. For three of these contingents (06, 10, and 11), the mean for all first-time examinees was high

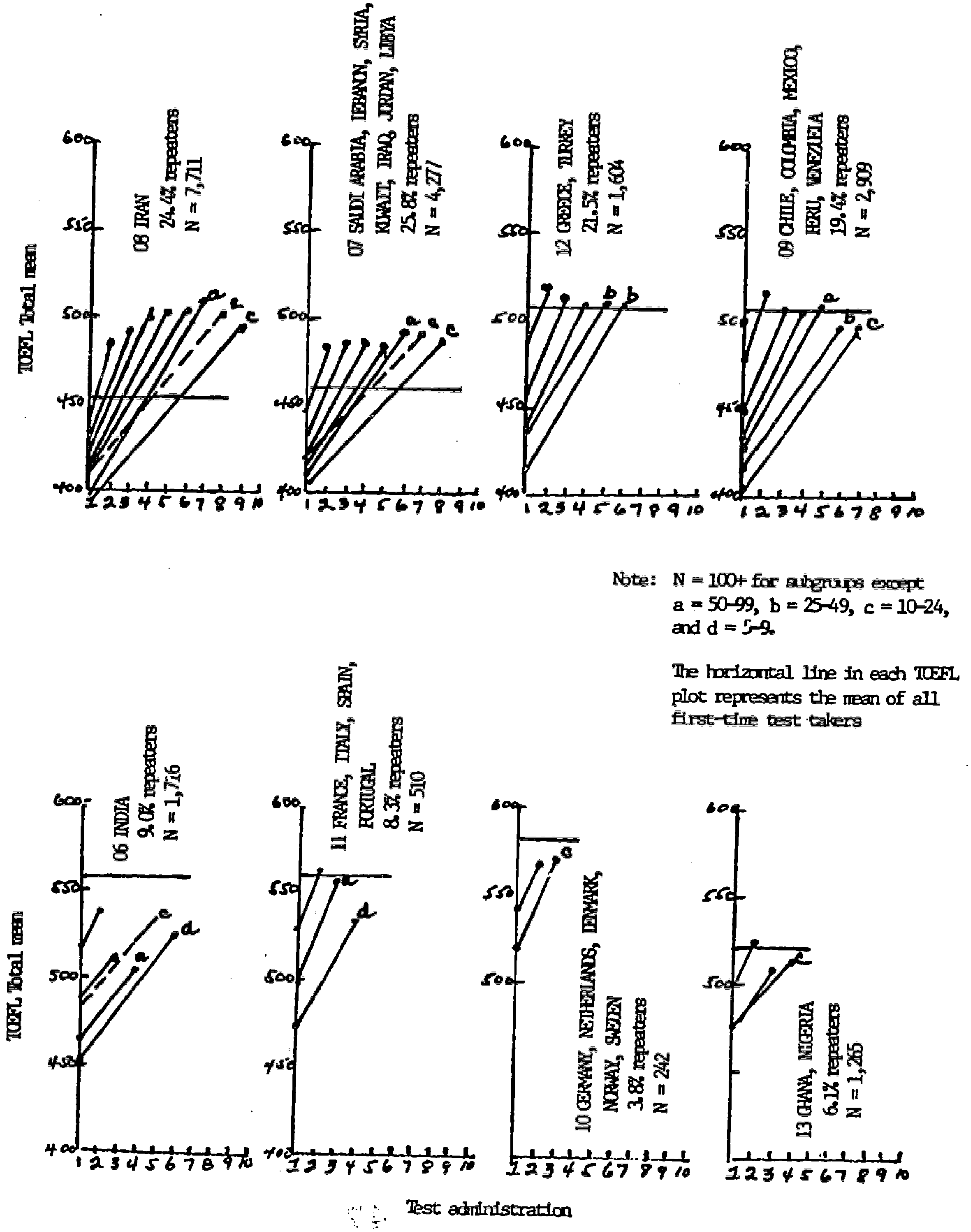
Figure 2. Plot of initial and final TOEFL total means for 2-, 3-, . . ., 10-time test takers, by analysis group, compared with plots of initial and most recent GRE verbal test means for 2-, 3-, 4-, and 5-time test takers



Note: N = 100+ for subgroups except a = 50-99, b = 25-49, c = 10-24, and d = 5-9.

The horizontal line in each TOEFL plot represents the mean of all first-time test-takers

Figure 2, Page 2



Note: N = 100+ for subgroups except a = 50-99, b = 25-49, c = 10-24, and d = 5-9.

The horizontal line in each TOEFL plot represents the mean of all first-time test takers (over 550), and repeaters in the fourth low-incidence group (13), and in group 06, may have had substantial exposure to academic instruction in English. Thus, at time of initial TOEFL-taking, examinees in the low- incidence contingents may have reached a more advanced level of maturation with respect to the development of English proficiency than examinees in the high-incidence contingents.

Differences in patterns of multiple test taking. There were relatively sharp differences, by analysis group, in the behavior of repeater subgroups with comparable t-1 means. More specifically, there were differences in the persistence (reflected in number of times tested) of repeaters across analysis groups, relative to average score gains and initial score levels. Figure 3 shows, illustratively, the means for successive test administrations for 2-, 3-, . . ., t-time test takers in four analysis groups, two with high and two with moderate overall incidence of test repetition.

In each of the four analysis groups, a repeater subgroup with a first-time mean of approximately 450 (a "t-1 450" subgroup) may be identified. In contingent 07, the "t-1 450" subgroup took only one additional test and earned a t-2 mean approaching 485; in contingent 09 this subgroup repeated TOEFL twice and earned a t-3 mean exceeding 500, and in contingents 01 and 02 the "t-1 450" subgroups were made up of 8-time and 10-time test takers, with t-8 and t-10 means of approximately 490 and 520.

It seems reasonable to believe that the "t-1 450" subgroup in group 07, which averaged around 485 on the final testing, might have attained a mean approaching 500 on a third administration of TOEFL after additional effort. However, they did not persist. On the other hand, the "t-1 450" repeater subgroup in the Hong Kong contingent, which averaged approximately 485 at t-3, persisted through 10 test administrations. The reasons for such differences in test-taking behavior are not evident. None of the times-tested subgroups in analysis group 07 (Arabic Mideast) persisted through attainment of t-T means of 500 despite mean gain patterns suggesting the feasibility of attaining this goal with moderate additional effort. At the same time, repeaters with comparable t-1 means in other analysis groups persisted through eight to 10 testings and earned t-T means approaching or exceeding 500.

Such a pattern may reflect differences in perceived accomplishment--the final performance of 2-, 3-, . . ., t-time test takers in analysis group 07 was quite high relative to the mean for all first-time test-takers and for that reason may have been perceived as sufficient, while in analysis groups 01 and 02, with considerably higher time-1 means, the score level associated with comparable gains may have been perceived as insufficient. Additional study is needed to shed light on the dynamics underlying the differences in test-taking patterns that are characteristic of the various contingents of examinees.

These findings support the proposition that in a population of U.S.-bound foreign examinees for whom English is a second language, TOEFL is measuring developing English language verbal skills, and that in the largely native English-speak- ing GRE population, the verbal abilities measured by the GRE verbal measure are more fully developed, thus less amenable to change.

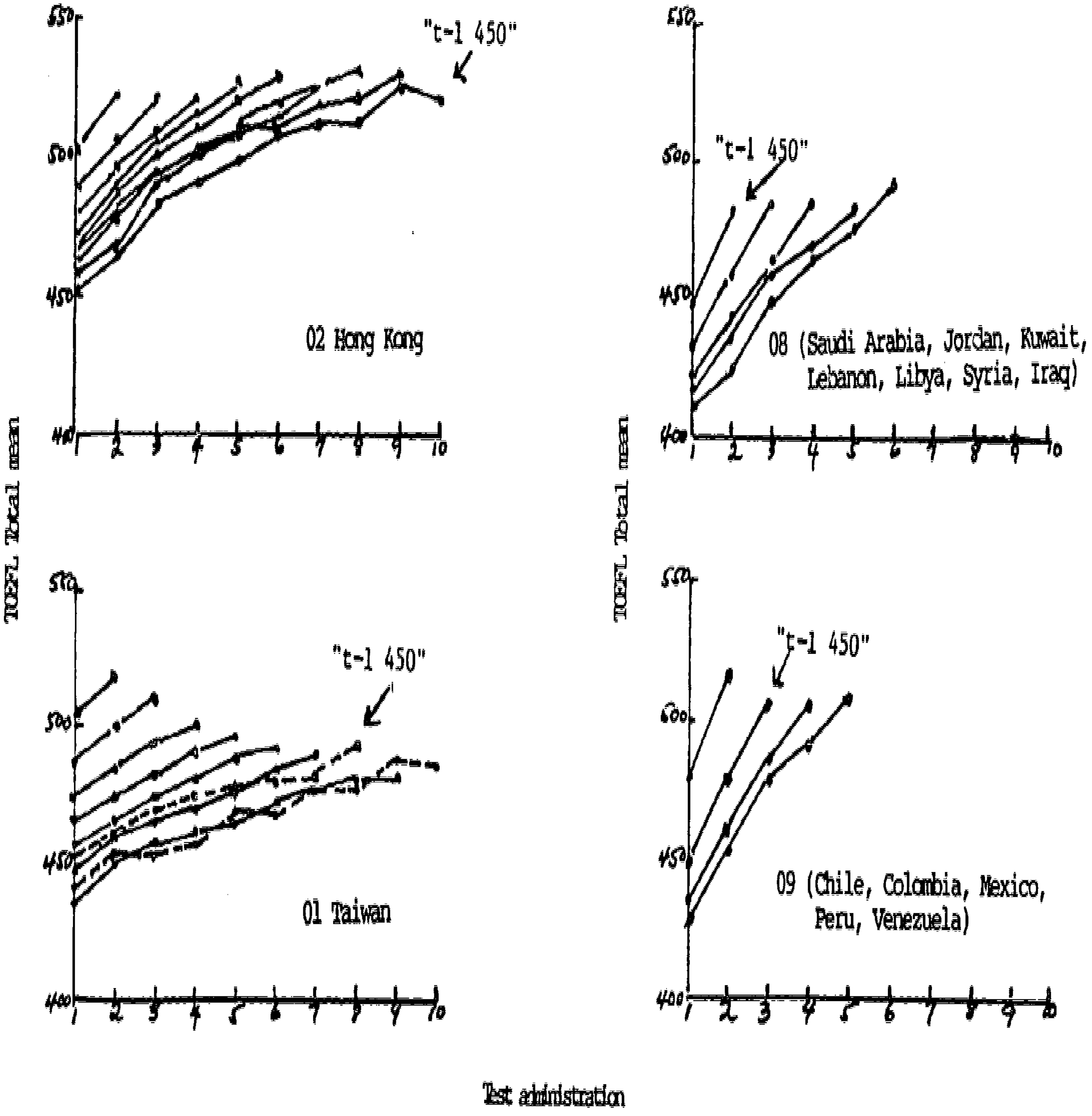


Figure 3. Means on successive TOEFL administrations of 2-, 3-, . . ., t-time test takers in four analysis groups

ERIC

Section VI. Mean Change in TOEFL Section Scores

Tables 8 and 9 summarize data on the t-1 and t-T section score means, and the corresponding differences in means, by analysis group. In Table 9, data are presented for analysis groups ordered from high to low in terms of mean change on TOEFL total. Ranks of the groups on the section-score mean change values are also shown.

For eight of the analysis groups, the lowest means were obtained on Structure and Written Expression (SWE); for only one group was this mean the highest. For six groups, mean Listening Comprehension (LC) scores were highest; means on Reading Comprehension and Vocabulary (RC&V) were highest for six additional groups. Of the six groups with the highest mean total score gain, five were among the six highest gainers on RC&V and SWE; all had LC-gain ranks between 1 and 6, out of 13.

Across analysis groups, ranks with respect to mean SWE gain corresponded a bit more closely to ranks on total gain (rho = .962) than did ranks on LC (rho = .924) or RC&V (rho = .902).

Figure 4 shows comparative profiles of t-1 and t-T section score means for repeaters, and comparable profiles of means for nonrepeaters, in each analysis group. In five analysis groups (02, 04, 05, 07, and 08), the final (t-T) means of repeaters (broken lines with dots) exceeded those for nonrepeaters (solid lines with x's) on all three sections. In four groups (06, 09, 10, and 11) the opposite was true, and in the remaining four groups (01, 03, 12, and 13) the t-T means of repeaters were lower than those of nonrepeaters on at least one of the sections.

Except for the Korean repeaters, the t-1 and t-T mean profiles of repeaters (solid lines with dots) and the profiles of nonrepeaters were roughly parallel—nonrepeaters in group 03 (Korea) had an atypically high mean on Listening Comprehension, and the t-T LC mean of repeaters was slightly lower than that of nonrepeaters, although on both SWE and RC&V, the repeater t-T means surpassed the means of nonrepeaters.

Similarities and differences among the analysis groups with respect to relative performance on the respective TOEFL sections are discernible at both t-1 and t-T, reflecting similarities and differences in the relative development of the corresponding English language skills. Repeaters from Taiwan, Hong Kong, Thailand, and Japan (groups 01, 02, 04, and 05), for example, had generally similar profiles. Profiles for Indian and Korean repeaters (groups 03 and 06) differ from the others, being characterized by relatively low means on LC, with progressively higher means on SWE and RC&V. Unique among the analysis groups, the profiles of repeaters from Ghana and Nigeria (group 13) exhibited considerably lower LC than SWE means, and somewhat higher SWE than RC&V means.

Profiles for the two Mideastern analysis groups (07 and 08) are generally similar, characterized by substantially higher LC than SWE means, and somewhat higher SWE than RC&V means. Similar statements might be made about the other

Table 8

Mean First- and Last-Administration Section and Total Scores, by Analysis Group

Analysis group	First test administration				Last test administration			
	LC*	SWE*	RC&V*	Tot	LC	SWE	RC&V	Tot
01	49.1	48.1	49.7	489.9	50.9	49.9	51.8	508.9
02	49.2	48.2	49.8	490.6	52.6	51.3	52.6	521.8
03	47.7	49.4	52.0	496.9	50.2	52.2	54.3	522.3
04	47.2	44.9	45.4	458.0	50.3	47.9	48.4	488.6
05	47.9	46.6	46.5	470.1	51.5	49.9	50.0	504.5
06	48.8	50.6	53.0	508.0	51.9	53.2	54.9	533.2
07	47.7	42.4	41.8	439.7	52.1	47.0	46.1	484.0
08	47.6	42.0	40.8	434.8	53.2	47.1	46.2	488.4
09	48.6	43.1	48.2	465.9	53.0	48.7	52.2	512.8
10	57.6	51.5	52.4	538.3	60.0	55.1	55.2	567.5
11	52.2	50.4	54.4	523.2	56.2	54.4	57.1	559.2
12	50.1	45.9	46.0	473.4	54.0	50.5	50.0	515.4
13	46.2	51.2	51.1	494.5	49.6	54.1	53.1	522.4
Median	48.6	48.1	49.7	489.9	52.1	50.5	52.2	515.4

Note: Analysis groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudi Arabia, Kuwait, Jordan, Lebanon, Libya, Syria, Iraq), 08 (Iran), 09 (Venezuela, Mexico, Colombia, Peru, Chile), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Spain, Italy, Portugal), 12 (Greece, Turkey), 13 (Ghana, Nigeria).

* LC (Listening Comprehension); SWE (Structure and Written Expression); RC&V (Reading Comprehension and Vocabulary).

Table 9

Mean Change, First to Last Test Administration, in TOEFL Section Scores, by Analysis Group

Group	N	Mean change						
		LC*	Rank	SWE*	Rank	RC&V*	Rank	Total
08	7704	(5.6)	1	5.1	2	5.6	1	53.6
09	2906	(4.4)	2.5	5.6	1	4.0	3.5	46.9
07	4268	(4.4)	2.5	4.6	3.5	4.3	2	44.3
12	1604	(3.9)	5	4.6	3.5	4.0	3.5	42.0
11	510	4.0	4	4.0	5	(2.7)	9	36.0
05	7057	(3.6)	6	3.3	7	3.5	6	34.4
02	9612	3.4	7.5	3.1	8	(3.8)	5	31.2
04	3823	3.1	9.5	3.0	9	(3.0)	7	30.6
10	242	(2.4)	12	3.5	6	2.8	8	29.2
13	1265	3.4	7.5	(2.9)	10	2.0	12	27.9
03	4317	2.5	11	2.8	11	(2.3)	10	25.6
06	1715	3.1	9.5	2.6	12	(1.9)	13	25.2
01	16027	1.8	13	1.8	13	(2.1)	11	19.0

Note: Groups, listed in descending order of total score gain, are 08 (Iran), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 07 (Saudi Arabia, Kuwait, Lebanon, Jordan, Syria, Liby, Iraq), 12 (Greece, Turkey), 11 (France, Italy, Spain, Portugal), 05 (Japan), 02 (Hong Kong), 04 (Thailand), 10 (Germany, Netherlands, Denmark, Sweden, Norway), 13 (Nigeria, Ghana), 03 (Korea), 06 (India), 01 (Taiwan).

* LC (Listening Comprehension); SWE (Structure and Written Expression); RC&V (Reading Comprehension and Vocabulary). Underscoring indicates the section on which the repeaters in each analysis group had the lowest t-1 mean; the section with the highest mean is indicated by parentheses.

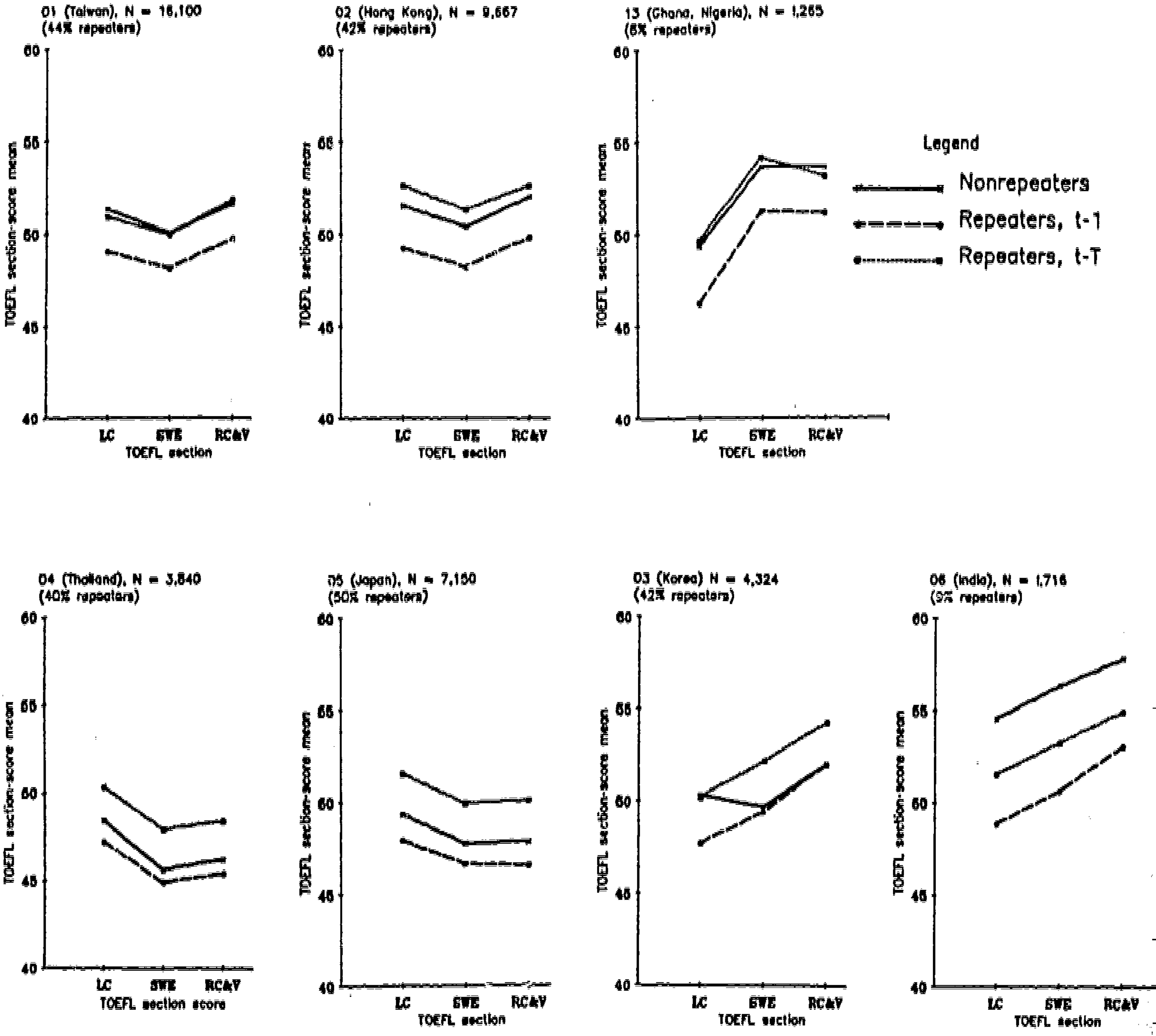
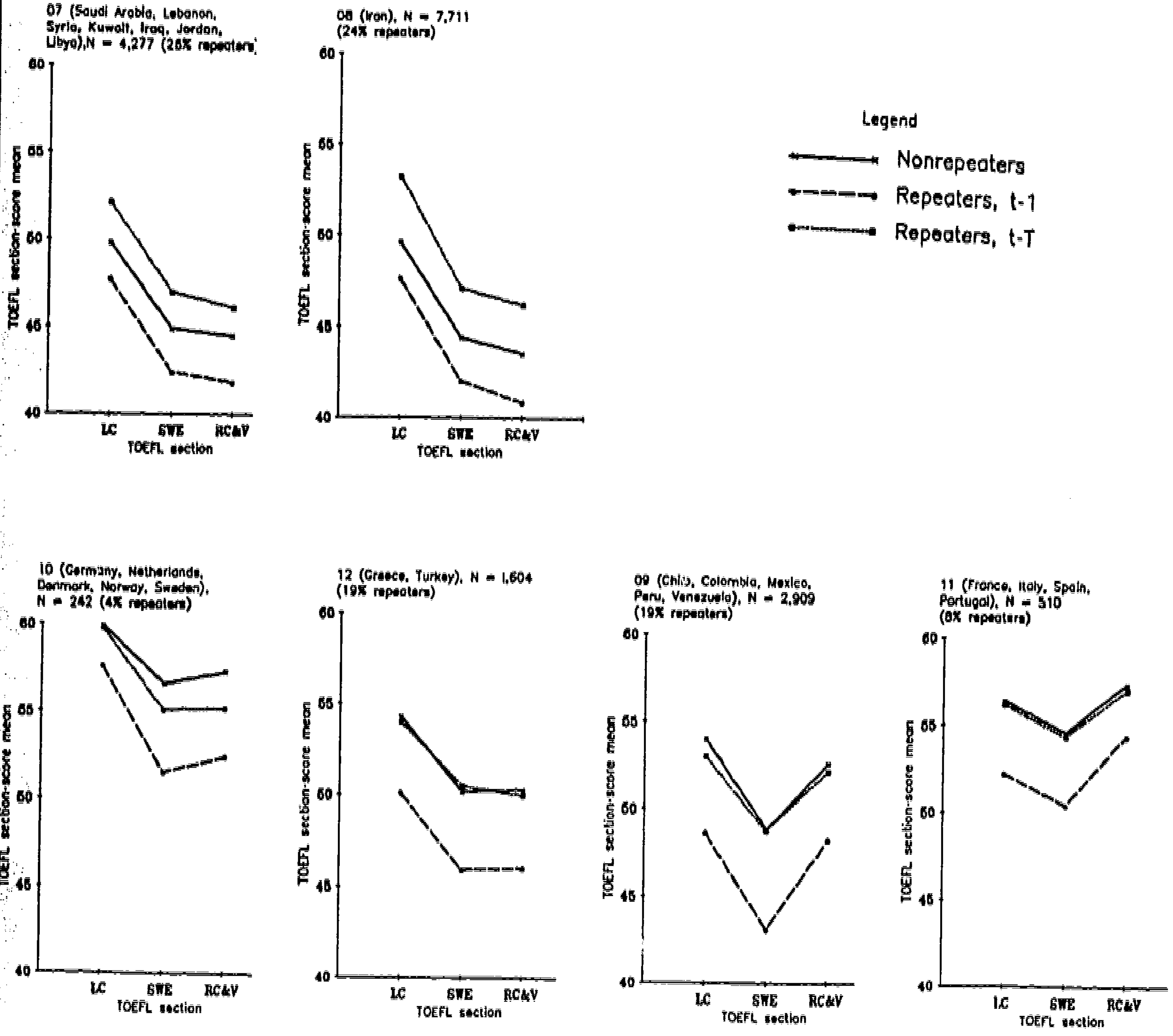


Figure 4. First-time (t-1) and final (t-T) TOEFL section-score means for repeaters relative to section-score means for nonrepeaters, by analysis group



07 (Saudi Arabia, Lebanon, Syria, Kuwait, Iraq, Jordan, Libya),N = 4,277 (26% repeaters)

08 (Iran), N = 7,711 (24% repeaters)

10 (Germany, Netherlands, Denmark, Norway, Sweden), N = 242 (4% repeaters)

12 (Greece, Turkey), N = 1,604 (19% repeaters)

09 (Chile, Colombia, Mexico, Peru, Venezuela), N = 2,909 (19% repeaters)

11 (France, Italy, Spain, Portugal), N = 510 (8% repeaters)

four groups. Perhaps the most pertinent generalization is that the development of the components of English proficiency represented by the TOEFL sections seems to remain relatively consistent across test repetitions. For each analysis group there are clear similarities in the patterning of t-1 and t-T section-score mean profiles of repeaters; with the exception of the Korean contingent, the profiles of repeaters and nonrepeaters for the respective analysis group are similar in shape.

Section VII. Selected Correlates of Change in TOEFL Total Score

The preceding sections have been concerned primarily with patterns of TOEFL-taking and average change in TOEFL scores for the analysis groups. This section discusses (a) selected variables as correlates of change in TOEFL scores, and (b) the relative contribution of those variables to prediction of final (t-T) total score. Unless otherwise noted, all the analyses reported are based on a 20 percent random sample of examinees (repeater N = 11,612).

The variables selected for study were as follows:

Variable	Definition
Test variables	
Change or Difference (D)	Difference between the last score of record (time t or t-T) and the initial score (time 1 or t-1)
LC-D	Listening Comprehension change
SWE-D	Structure & Written Expression change
RC&V-D	Reading Comprehension & Vocabulary change
Total-D	TOEFL total change
LC-1 SWE-1 RC&V-1 TOTAL-1	First-time (t-1) score on variable
LC-t SWE-t RC&V-t Total-t	Last score of record (t-T) for variable
Nontest variables	
Intrvl	Interval in months between t-1 and t-T
Numtsts	Number of times tested (2, 3, . . ., 10)
Edlevl	At t-1, graduate = 1; undergraduate = 0
Sex	Male = 1, Female = 0
Center-U.S.	At t-T, U.S. center = 1; Other = 0
Reports	At t-1, designated institution to receive score report = 1; did not do so = 0
Age	At t-1, age in years

Table 10 shows the means of the 13 analysis groups, and of the total 20 percent sample of repeaters, on the non-test variables. The follow- ing general picture of TOEFL repeaters and their behavior emerges:

Table 10

Means on Selected Variables for Repeaters, by Analysis Group

		Variables*						
Group	N**	Intrvl	Numtsts	Edlevl	Sex	Center	Reports	Age
01	3070	11.6	2.9	0.83	1.41	0.08	0.34	25.1
02	1781	13.7	3.2	0.26	1.34	0.06	0.24	19.3
03	819	11.7	2.8	0.76	1.26	0.26	0.38	25.3
04	733	12.9	3.0	0.85	1.44	0.32	0.34	23.8
05	1346	11.3	3.2	0.50	1.34	0.40	0.41	23.8
07	814	12.6	2.7	0.34	1.07	0.65	0.35	21.8
08	1500	9.9	2.7	0.31	1.21	0.79	0.33	21.3
09	566	10.1	2.5	0.60	1.26	0.69	0.46	23.9
12	307	10.2	2.6	0.56	1.16	0.34	0.45	21.4
06	313	15.5	2.3	0.82	1.08	0.18	0.72	23.3
10	48	11.5	2.1	0.46	1.27	0.15	0.46	22.5
11	101	10.6	2.2	0.78	1.24	0.22	0.57	24.0
13	214	15.9	2.1	0.06	1.09	0.22	0.57	22.2
Total	11612	11.9	2.9	0.57	1.30	0.32	0.36	23.0

Note: Groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 07 (Saudi Arabia, Kuwait,Jordan, Lebanon, Iraq, Syria), 08 (Iran), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 12 (Greece, Turkey), 06 India, 10 Germany, Netherlands, Denmark, Norway, Sweden), 11 France, Italy, Spain, Portugal), 13 (Ghana, Nigeria).

* Interval in months between T1 and Tt; number of times tested; educational level (graduate = 1, undergraduate = 0); sex (M = 1, F = 2); center (last test in U.S. = 1, not in U.S. = 0); reports (1 = designated institution to receive t-1 report, 0 = did not do so); age in years at time initial testing (t-1).

** Analyses are based on a 20 percent random sample of all repeaters.

o On the average, these TOEFL repeaters were tested (in International and Special Center administrations) approximately three times (mean Numtsts = 2.9); the average interval in months between the first test administration and the last of record was about one year (11.9 months).

o Some 30 percent of the repeaters were females (mean sex = 0.30); average age at t-1 was 23 years. More than one half of the repeaters (57 percent) planned graduate study in the United States or Canada at time of initial testing (Edlevl = 0.57), but only slightly more than one-third designated institutions to receive their t-1 TOEFL score report (mean reports = 0.36).

o Approximately one third (32 percent) were in the United States when they took the last test of record (mean Center-U.S. = 0.32). Data not tabled indicate that about six in 10 of this subgroup of examinees took both t-1 and t-T in the United States.

o There were differences among the 13 analysis groups with respect to each of the nontest variables.

Means, standard deviations, and intercorrelations of the test and nontest variables in the 20 percent sample of repeaters, undifferentiated with regard to analysis group, are shown in Table 11. The upper portion of the table shows intercorrelations of the 12 test variables; the middle portion (rows 13 through 19) show the simple correlation of nontest variables with the 12 test variables; and the bottom portion shows the intercorrelations of the seven nontest variables.

The coefficients in row 4 indicate, for example, the correlation of change between t-1 and t-T in TOEFL total score (TOT-D) with each of the other test variables. The coefficients in row 13, for example, indicate that time interval between t-1 and t-T was positively related to change in TOEFL section and total scores, negligibly related to t-1 scores, and positively related to final (T) scores. It may also be determined from the middle rows that change in TOEFL total (TOT-D) was negatively related to Age (r = -.13), Edlevl (r = -.12, and Sex (r = -.08), indicating somehat greater average net gains for younger repeaters, undergraduates, and males; that repeaters who designated institutions to receive their initial score reports had slightly lower average gain (r = -.07) than their counterparts who did not have their reports sent to institutions; and so on.

In evaluating the correlations between the initial scores on TOEFL and the difference or change between t-1 and t-T, it is important to keep in mind that these coefficients do not reflect the relationship between "true initial standing" and "true gain," since TOEFL is not a perfectly reliable measure of its underlying construct. Lord (1967, pp. 33-34) has provided formulas for estimating the relationship between true initial standing and true gain. For the data of this study, assuming a reliability of .9 for both t-1 and t-T, the estimated correlation for true t-1 total score standing and total score gain is -.39, very close to the observed coefficient (r = -.43). The similarity between these two coefficients supports the construct validity of TOEFL.

Table 11

Means, Standard Deviations, and Intercorrelatons of Variables for Repeaters

Intercorrelations of test variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12
	LC	SWE	RC&V	TOT	LC	SWE	RC&V	TOT	LC	SWE	RC&V	TOT
	D	D	D	D	1	1	1	1	T	T	T	T
1 LC-D	1.00	.30	.32	.72	-.47	-.17	-.19	-.32	.46	.08	.05	.23
2 SWE-D		1.00	.37	.75	-.08	-.47	-.18	-.29	.19	.39	.10	.27
3 RC&V-D			1.00	.74	-.13	-.23	-.50	-.35	.16	.08	.27	.20
4 TOT-D				1.00	-.30	-.39	-.38	-.43	.37	.25	.18	.32
5 LC-1					1.00	.45	.46	.73	.56	.40	.40	.54
6 SWE-1						1.00	.72	.88	.29	.63	.61	.62
7 RC&V-1							1.00	.88	.28	.60	.70	.64
8 TOT-1								1.00	.44	.66	.70	.72
9 LC-T									1.00	.48	.45	.75
10 SWE-T										1.00	.73	.88
11 RC&V-T											1.00	.87
12 TOT-T												1.00
Mean	3.3	3.2	3.1	32.1	48.4	46.4	47.5	474.6	51.7	49.7	50.6	506.6
S.D.	5.7	5.8	5.4	41.4	6.1	6.9	7.3	56.5	6.1	6.6	6.5	53.8

Correlation of selected characteristics with test variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12
13 Interval	.13	.17	.17	.22	-.03	.01	.01	-.00	.09	.17	.15	.16
14 Numtsts	.19	.18	.19	.25	-.19	-.16	-.16	-.20	-.02	-.01	-.02	-.02
15 Edlevel	-.11	-.09	-.08	-.12	-.01	.16	.22	.16	-.11	.09	.18	.07
16 Sex	-.08	-.08	-.04	-.08	.08	.06	-.00	.05	.01	-.00	-.03	-.01
17 Center	.24	.19	.20	.28	-.09	-.33	-.38	-.33	.13	-.18	-.26	-.13
18 Reports	-.04	-.05	-.07	-.07	.09	.13	.13	.14	.05	.09	.09	.09
19 Age	-.10	-.08	-.10	-.13	-.14	.02	.12	-.01	-.24	-.05	.05	-.09

Intercorrelations of characteristics

	Inter-val	Num-tsts	Edlev	Sex	Cen ter	Re-ports	Age	Mean	S.D.
13 Interval	1.00	.39	-.02	.00	.05	-.03	-.09	11.88	9.21
14 Numtsts		1.00	.01	-.01	.04	-.11	.03	2.87	1.39
15 Edlevel			1.00	-.01	-.22	.00	.44	0.57	0.50
16 Sex				1.00	-.06	-.04	-.10	1.30	0.46
17 Center					1.00	.03	-.08	0.32	0.47
18 Reports						1.00	.04	0.36	0.48
19 Age							1.00	22.99	4.45

Note: These summary statistics are for a 20 percent spaced sample of repeater-records, N = 11,612. D = difference between final (t-T) and initial (t-1) scores (change); 1 = initial, t-1 score, and T = final, t-T score; **Interval** = t-1 to t-T interval in months,; **Numtsts** = number of times tested (2, 3, . . , 10); **Edlevel** (graduate = 1, undergraduate = 0); **Sex** (M = 1, F = 2); **Center** (1 = took last test in U.S. center, 0 = elsewhere); **Reports** (1 = designated schools to receive score reports, 0 = did not do so); **Age** (in years at time of t-1).

The circumstances (correlated errors of measurement) leading to elements of spuriousness in correlations between initial standing and gain based on two less than perfectly reliable tests (or other measures) are not involved when other variables are correlated with change (Lord, 1967). Thus, for example, the correlations of variables such as Interval between t-1 and T-T and number of times tested with change, or with t-1 or t-T TOEFL scores, are not affected by correlated errors of measurement. Accordingly, when the observed correlation between t-1 score and change is introduced to control for initial standing on TOEFL for purposes of assessing the relative contribution of nontest variables to change, correlated errors of measurement are not involved (see following section).*

Stability of TOEFL scores. Coefficients reflecting the stability of relative standing, T-1 to t-T, of examinees in terms of performance on LC, SWE, RC&V, and Total, are shown in the submatrix specified by columns 9-12, rows 5-8: T-1 to t-T stability coefficients were .56, .63, .70, and .72 for the respective measures. Listening comprehension was the least stable, and reading comprehension and vocabulary was the most stable of the English-language skills measured by the TOEFL.

The stability coefficient (.72) for TOEFL total, over periods averaging approximately one year, may be compared with the internal consistency reliability coefficients of approximately .9 reported for this score (e.g., Educational Testing Service, 1983), and a coefficient of .92, reported by Swinton (1983), for scores of 98 students in an intensive language program on two forms of TOEFL administered one week apart. The coefficient is also lower than comparable test-retest stability coefficients involving verbal admission tests in samples of U.S. examinees (see, for example, Alderman, 1981; Rock and Werts, 1979; Wilson, 1983, Table 4). The distinction between "developing English language skills" in samples of ESL examinees and "developed verbal reasoning abilities" in samples of U.S. examinees is pointed up by both (a) the lower stability coefficients for TOEFL in samples of ESL examinees, than for English language verbal admission tests in samples composed primarily of native speakers and (b) the differences in patterns of score change for TOEFL and GRE repeaters (shown earlier in Figure 2).

The time-1 to time-T stability coefficient reflects the t-1/t-T correlation in a sample composed of 2-, 3-, . . . , 10-time test takers, for whom the t-1 to t-T interval varied markedly, averaging about 12 months. Trends in t-1/t-T correlation for 2, 3, . . . , 9-time test takers (the Numtsts subgroups with N = 50+ in the sample), as well as total score means and standard deviations for each test administration, are shown in Table 12. The overall stability coefficients, t-1 to t-T, tend to decrease as Numtsts increases—that is, stability tends to decline as average amount of change increases (and as the average t-1 to t-T time interval increases), though with some reversals in the declining trend. The data in Table 12

* The writer is grateful to Brent Bridgeman for calling attention to these important points.

Table 12

Trends in Correlation Between Time-1 and Time-T Total Scores, by Number of Times Tested, for Repeaters Without Regard to Analysis Group

Correlation between t-1 score and score for designated administration

Times tested	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	(N)
(2)	1.000	.784 (9.0)*								6,638
(3)	1.000	.800	.674 (13.4)							2,577
(4)	1.000	.788	.719	.590 (15.9)						1,128
(5)	1.000	.768	.732	.718	.574 (18.2)					608
(6)	1.000	.750	.769	.697	.644	.586 (20.5)				307
(7)	1.000	.803	.750	.730	.688	.666	.594 (20.7)			175
(8)	1.000	.762	.714	.700	.687	.691	.632	.645 (22.9)		104
(9)	1.000	.748	.767	.692	.757	.705	.739	.684	.552 (24.1)	51

* Mean interval in months between designated administrations, t-1 to t-T.

Means for designated administrations

	1	2	3	4	5	6	7	8	9
(2)	483.8	507.5							
(3)	468.8	488.4	505.6						
(4)	459.9	477.4	490.5	505.5					
(5)	452.5	467.6	479.4	491.2	503.6				
(6)	452.1	468.0	480.1	488.3	498.2	506.2			
(7)	446.3	462.2	472.6	480.4	487.2	492.7	500.9		
(8)	445.7	462.2	472.6	480.9	485.5	492.3	495.7	509.5	
(9)	440.5	457.6	463.4	472.8	478.8	483.9	488.8	495.2	501.9

Standard deviations

	1	2	3	4	5	6	7	8	9
(2)	58.8	57.8							
(3)	52.3	49.2	49.6						
(4)	49.2	45.0	45.8	46.7					
(5)	46.8	44.0	44.3	45.9	46.9				
(6)	47.3	41.0	42.0	40.1	40.9	42.1			
(7)	45.8	43.1	44.6	45.6	44.9	45.3	47.7		
(8)	44.1	42.7	39.2	40.0	39.1	40.7	41.5	49.9	
(9)	46.0	44.5	48.7	44.9	45.4	40.2	43.9	48.1	49.7

Note: Ns for means and standard deviations are as shown for correlations. These analyses are based on a 20 percent sample of repeaters.

attest further to the developing nature of the skills measured by the TOEFL in samples of ESL examinees.

Relative Contribution of Nontest Variables to Prediction of Change

In order to assess the relative contribution of the seven nontest variables to the prediction of change in TOEFL total within each of the 13 analysis groups and in the total sample, two multiple regression analyses were conducted. In the first analysis, change in TOEFL total (TOT-D) was regressed on the seven nontest variables only; in the second analysis, the initial TOEFL total score was added to the set of nontest variables. Results of the two analyses are provided in Table 13. The first row for each group shows standard partial regression coefficients for the seven non-test variables only, and the corresponding multiple correlation coefficients; the second row shows comparable data for the analysis in which t-1 TOEFL total score was added to the battery. The four non-test variables with the highest partial regression (beta) weights are indicated by bold-face type.

The general-sample results in Table 13 indicate that the same set of four variables (**Numtsts, Intrvl, Center-U.S., and Age**) contributed most to prediction of **change** in TOEFL total in both the analysis without control for t-1 score and that with control for t-1 score. However, introduction of control for t-1 score resulted in changes in the relative magnitudes of the regression weights for these variables: without control for t-1 score, **Numtsts** and **Center-U.S** are more heavily weighted than **Intrvl** and **Age**, but with control for T-1 score, **Center-U.S.** has a slightly lower weight than **Intrvl; Numtsts** and **Age** contribute about equally to prediction, with slightly lower weights than those for the **Intrvl** and **Center-U.S.** Sex (M = 1, F = 2) was negatively weighted in the total sample analysis, as was t-1 **score reporting (Reports).**

With some variation in detail, the pattern of findings in the respective analysis groups was similar to the pattern described above: with control for the t-1 score, **Intrvl** was more heavily weighted than either **Numtsts** or **Center-U.S** in 10 of 13 groups, but without control for t-1 score, either **Numtsts** or **Center-U.S.** was more heavily weighted than **Intrvl** in 11 groups. **Sex** was negatively weighted in ten of the analysis groups, with control over t-1 score, and the findings for **Reports** were similar.

Relative Contribution of Variables to Prediction of t-T Total Score

Table 14 shows the results of regression analyses in which the dependent variable was the last TOEFL total score of record (the t-T score), and the independent variables were the t-1 score and the seven non-test variables. Simple **t-1/t-T** correlations and multiple correlation coefficients are shown, along with standard partial regression weights for various independent variables. As expected from the preceding analysis (in which change, **t-1 to t-T**, was the dependent variable), (a) inclusion of the nontest variables resulted in enhanced prediction of the t-T total score, and (b) **Age, Intrvl, Center,**

Table 13

Results of the Regression of Change (C = t -T - T-1) in TOEFL Total, on Selected Personal, Academic, and Testing-Related Variables, with and without Control for Time-1 Total Score, by Analysis Group

		Standard partial regression weights								Multiple
Group	N	1	2	3	4	5	6	7	8	correlation
		Total-1	Numtsts	Incr-vl	Center	Edlevel	Age	Sex	Reports	R
01	3070		.200*	.0[illegible]09	.081*	-.008	-.078	-.086	-.110*	.278
		-.340	.091*	.0[illegible]45	.070*	.060	-.124*	-.094	-.037	.410
02	1781		.302*	.1[illegible]6*	.100*	-.013	-.085*	-.054	-.048	.456
		-.269	.241*	.1[illegible]8*	.098*	-.037	-.103*	-.051	-.026	.524
03	819		.176*	.1[illegible]2*	.124*	.004	-.074*	-.062	-.023	.338
		-.315	.116*	.1[illegible]2*	.094*	.084	-.110*	-.093	-.012	.450
04	733		.156*	.1[illegible]2*	.213*	-.016	-.081	-.071	-.098*	.430
		-.302	.105*	.1[illegible]2*	.178*	.015	-.117*	-.040	-.047	.517
05	1346		.262*	.2[illegible]6*	.213*	-.026	-.104*	.021	-.051	.532
		-.275	.207*	.2[illegible]9*	.133*	.006	-.062*	.023	-.030	.585
06	313		.121*	.0[illegible]7	.142*	.069	-.061*	-.033	.068*	.205
		-.448	.059	.1[illegible]7*	.086	.111*	-.130*	.033	.100*	.446
07	814		.179*	.1[illegible]6*	.126*	.001	-.056*	-.036	-.010	.356
		-.319	.110*	.20[illegible]4*	.078*	.026	-.108*	-.016	.035	.462
08	1500		.208*	.21 4*	.010	.002	-.026*	-.004	-.079*	.364
		-.420	.137*	.22[illegible]4*	-.042*	.009	-.107*	-.008	-.024	.540
09	566		.251*	.21 5*	.215*	.054	-.159*	-.011	-.013	.434
		-.464	.114*	.29 8*	.133*	.133*	-.212*	-.016	.023	.601
10	48		.165*	.09 8*	.259*	.010	-.182*	-.098	-.126	.382
		-.237	.166*	.22 1*	.257*	-.021	-.117	-.130	-.155*	.430
11	101		-.081*	.16 1*	.475*	.012	.103*	-.066	-.032	.503
		-.392	-.098*	.19 7*	.396*	.098*	.016	-.026	-.026	.622
12	307		.339*	.26[illegible]1*	.218*	-.014	.069	-.094	-.088	.572
		-.363	.215*	.26[illegible]4*	.124*	.040	.006	-.055	.024	.649
13	214		.205*	-.11[illegible]*	.048	.033	-.106*	-.057	-.043	.246
		-.396	.107*	-.03[illegible]6*	.047*	-.030	-.102*	.000	.027	.445
Total	11612		.190*	.116[illegible]	.253*	-.026	-.098*	-.076	-.056	.412
		-.359	.115*	.152[illegible]	.142*	.016	-.119*	-.065	-.008	.525

Note. Analysis groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 07 (Saudi Arabia, Kuwait, Jordan, Lebanon, Iraq, Syria), 08 (Iran), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 12 (Greece-Turkey) 06 (India), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Italy, Spain, Portugal), 13 (Ghana-Nigeria). * = highest regression weight[illegible] for nontest variables.

Table 11

Regression of t-T TOEFL Total Score on t-1 Score and Selected Testing-Related, Academic, and Demographic Variables, by Analysis Group

		Standard partial regression weights								R	r
		1	2	3	4	5	6	7	8		
Group	N	Total	Numtsts	Intrvl	Center	Edlevl	Age	Sex	Reports	tT.(1-8)	(t1,tT)
01	3070	.758	.065	.032	.050	.043	-.088	-.067	-.026*	.760	.746
02	1781	.768	.181	.134	.074	-.028*	-.077	-.038	-.020*	.769	.710
03	819	.800	.075	.118	.061	.055	-.072	-.060	-.008*	.814	.790
04	733	.749	.077	.134	.131	.011*	-.086	-.029	-.034	.777	.723
05	1346	.885	.137	.165	.088	.004*	-.041	.015*	-.020*	.844	.797
07	814	.681	.088	.164	.063	.021*	-.087	-.012*	-.028*	.701	.657
08	1500	.528	.130	.213	-.040	.008*	-.102	-.008*	-.022*	.602	.517
09	566	.702	.089	.234	.105	.105	-.166	-.012*	.018*	.778	.712
12	307	.847	.174	.213	.100	.032*	.004*	-.044	-.019*	.790	.723
06	313	.845	.030*	.079	.043	.066	-.065	.017*	.050	.893	.884
10	48	.773	.097	.130	.150	-.012*	-.068	-.076	-.091	.848	.816
11	101	.755	-.079	.159	.319	.079	.013*	-.021*	-.020*	.776	.700
13	214	.731	.082	-.028*	.036	.023*	-.079	.000*	.021*	.723	.714
Tot	11612	.775	.088	.117	.109	.013*	-.092	-.050	-.006*	.755	.719

Note: Analyses are based on a 20 percent random sample of all repeaters. Groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 07 (Saudi Arabia, Kuwait, Lebanon, Iraq, Jordan, Syria), 08 (Iran), 09 (Chile, Mexico, Colombia, Peru, Venezuela), 12 (Greece-Turkey), 06 (India), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Italy, Spain, Portugal), 13 (Ghana-Nigeria).

* Variable contributes less than .001 to R-squared in R (t-T.12345678).

and Numtsts were the major nontest contributors to prediction in the total sample and in most of the analysis groups.

Number of times tested may be an indirect measure of effort, motivation, financial resources, practice effects, and so on. A longer interval between testings means greater potential opportunity for study, practice, and general maturation of proficiency in English. Individuals with longer t-1 to t-T intervals are unlikely to be engaged in intensive English language study during the entire interval between the initial and the final testing. However, it is reasonable to assume that between the first and the last test adminstration most I&SC repeaters are motivated, in part, by a common general objective, namely, to improve their English proficiency as an element in their overall preparation for study in the United States. Precisely how they work toward this objective cannot be determined from the data under consideration. The findings indicate that both number of times tested and time interval between initial and final test administrations contribute independently to change in TOEFL score.

The positive contribution of being in the United States at time of final testing undoubtedly reflects, to some extent, the impact of the general language environment on language learning.

Section VIII. Observed vs Expected Final (t-T) TOEFL Total Score Means, by Analysis Group

Results of the total-sample regression analyses provide a basis for generating expected mean t-T scores for the respective analysis groups and determining the extent to which their observed t-T total score means are consistent with expectation. Using the total sample regression equation, which takes into account the t-1 score, Numtsts, Intrvl, Center, Edlevl, Sex, Age, and Reports, expected t-T TOEFL total scores (t-T exp) were computed for each of the 13 analysis groups. The equation was as follows:

$$\text{t-T exp} = .737\ (\text{T-1}) + 3.416\ (\text{Numtsts}) + .681\ (\text{Intrvl}) + 12.637\ (\text{Center}) + 1.369\ (\text{Edlevl}) - 1.107\ (\text{Age}) - 5.877\ (\text{Sex}) - .647\ (\text{Reports}) + 167.391.$$

For a composite based on this equation, R = .755; the standard error of estimate (for individual predicted t-T scores) is approximately 35 scaled score points.

Table 15 shows the actual t-1 and t-T means, the expected means, and the difference between the two (expected minus actual), by analysis group. Groups are ordered in terms of the algebraic difference (high positive to high negative) between actual and expected t-T means.

o Observed t-T means were 10 to 24 points higher than expected, based on the general sample regression equation, for groups 10 (Germany, Netherlands, Denmark, Norway, Sweden), 11 (France, Italy, Spain, Portugal), 12 (Greece, Turkey), and 09 (Chile, Colombia, Mexico, Peru, and Venezuela). Groups 10 and 11 had the two highest t-1 means, and were characterized by low incidence of repetition; groups 12 and 09 had comparatively low T-1 means, but were characterized by only moderate incidence of repetition. The Iranian and Korean repeaters had t-T means between 2 and 6 points higher than expected; the t-1 means for these two groups differed by about 60 points.

o Sample t-T means for Thailand and Taiwan were between 4 and 7 points lower than expected; smaller negative residuals are evident for repeaters from Hong Kong, Japan, Ghana-Nigeria, and the Arabic-speaking Mideastern countries (groups 02, 05, 13, and 07). For Indian repeaters (group 06), the expected mean and the actual mean were the same.

In evaluating these findings it should be kept in mind that the analysis was not controlled for "intervening experience" variables that might be expected to affect growth in English proficiency (e.g., amount, intensity, and quality of total English-language involvement between initial and final testing), with respect to which there may have been (may tend to be) average differences among examinees from countries making up the analysis groups under consideration.

The fact that European and South American repeaters had higher gains relative to expectancy than those from, say, the Mideast, Taiwan, or Japan may reflect differences in the amount, intensity, and/or quality of total

Table 15

Discrepancies Between Observed and Expected Time-T TOEFL Total Scores, by Analysis Group: Expected Score Based on An All-Repeater Regression Equation Employing Time-1 Score and Seven Nontest Variables*

Group	N	(1) Time 1 actual (mean)	(2) Time-T actual (mean)	(3) Time-T expected (mean)	Difference (2) - (3)
11	101	530.7	566.9	542.9	+ 24.0
09	566	465.8	513.9	501.5	+ 12.4
12	307	475.0	518.0	507.6	+ 10.4
10	48	531.7	555.7	544.1	+ 11.6
08	1500	434.4	488.6	483.0	+ 5.6
03	819	495.5	521.1	518.8	+ 2.3
06	313	501.2	526.0	526.0	+ 0.0
13	214	490.9	517.9	518.7	- 0.8
05	1346	470.8	503.1	504.2	- 1.1
02	1781	489.5	519.0	520.1	- 1.1
07	814	439.7	485.4	487.3	- 1.9
01	3070	490.7	508.8	512.7	- 3.9
04	733	454.4	485.1	492.5	- 7.4
Total	11612	474.6	506.8	506.8	0.0

Note: Analyses are based on a 20 percent sample of all repeaters. Groups, ordered in terms of actual minus expected means, are 11 (France, Italy, Portugal, Spain), 09 (Chile, Colombia, Mexico, Peru, Venezuela), 12 (Greece-Turkey), 10 (Germany, Netherlands, Denmark, Norway, Sweden), 08 (Saudi Arabia, Kuwait, Lebanon, Iraq, Jordan, Syria), 03 (Korea), 06 (India), 13 (Ghana-Nigeria), 05 (Japan), 07 (Iran), 02 (Hong Kong), 01 (Taiwan), 04 (Thailand).

* Expected t-T = .737 (T-1) + 3.416 (Numtsts) + .681 (Intrvl)
12.637 (Center) + 1.369 (Edlevl) - 1.107 (Age)
- 5.877 (Sex) - .647 (Reports) + 167.391.

English-language intervening experience that may be characteristic of contingents of examinees from these regions or countries who either elect, or are required, to repeat TOEFL. Such a pattern of findings may also reflect, to some extent, differences in the degree of "linguistic-distance" between English and the various languages involved, and associated differences in characteristic rates of acquisition of proficiency in English as a second language.

The potential value of living in an English-speaking environment as a factor influencing score gain is suggested by the finding that repeaters whose last test was taken in the United States gained over 12 points more, on the average, than those tested outside the United States with similar scores on other study variables.

Section IX. Tendency to Repeat as a Function of Initial Score Level

It was assumed at the outset that the tendency to repeat TOEFL is related inversely to initial score level. The findings that have been reviewed are generally consistent with that assumption. For example, among first-time test takers in all analysis groups the mean initial scores of those identified as repeaters were systematically lower than those who were identified as non-repeaters (see Figure 4); the four analysis groups with lowest percentages of repeaters were among the six with t-1 means above the median for all analysis groups, while only two of six groups with above-median percentages of repeaters were above the t-1 median (Table 4 and related discussion); and, within each analysis group, for subgroups classified by number of test repetitions, t-1 means tended to vary inversely with number of times tested (see Figure 2 and related discussion).

The analyses involved, however, did not examine the nature of the relationship between the tendency to repeat and initial score levels within the various analysis groups. Table 16 shows the percentage of repeaters among first-time test takers classified by TOEFL total score intervals within various analysis groups. For this analysis, certain of the 13 original analysis groups were consolidated: I = Hong Kong and Korea, originally 02 and 03; II = Taiwan, 01; III = Thailand and Japan, 04 and 05; IV = Saudi Arabia, Kuwait, Jordan, Lebanon, Iraq, Syria, and Iran, 07 and 08; V = Chile, Colombia, Mexico, Peru, Venezuela, and Greece-Turkey, 09 and 12; and VI = Germany, Netherlands, Denmark, Norway, Sweden (10), France, Italy, Spain, Portugal (11), Ghana- Nigeria (13), and India (06), combined.

Trends in the data are portrayed graphically in Figure 5, which plots the percentage of repeaters at successive score levels for each of the six analysis groups. The mean t-1 score for the repeaters is indicated.

o **It is quite clear that the tendency to repeat was not a simple linear inverse function of initial score level**—that is, the tendency to repeat did not increase (decrease) regularly as initial score level declined (increased) throughout the score range in any of the analysis groups. Rather, the tendency to repeat appeared to be inversely related to score level only in the upper score ranges. For example, in all groups between score-level 490 and score-level 590 (representing intervals 480-499 through 580-599) an inverse relationship tended to obtain. Across a larger intervening score range (roughly 490 down to 390), the tendency to repeat was relatively stable. And, percent repeating tended to vary directly with score level in the lower score ranges (below 390)—percentage repeating tended to decrease as score-level decreased across score intervals below 390.

o **It is also clear that the tendency to repeat was strongly associated with analysis group membership.** For example, except for examinees in the two lowest score intervals reported, Asian examinees (groups I, II, and III) were much more inclined to repeat than examinees in other groups; those from Thailand and Japan (group III) exhibited relatively strong tendencies to repeat at score levels above 540 and below 340. On the other hand, proportionately few

Table 16

Percentage of First-Time Test Takers Repeating TOEFL, by Initial Score-Level and Analysis Group

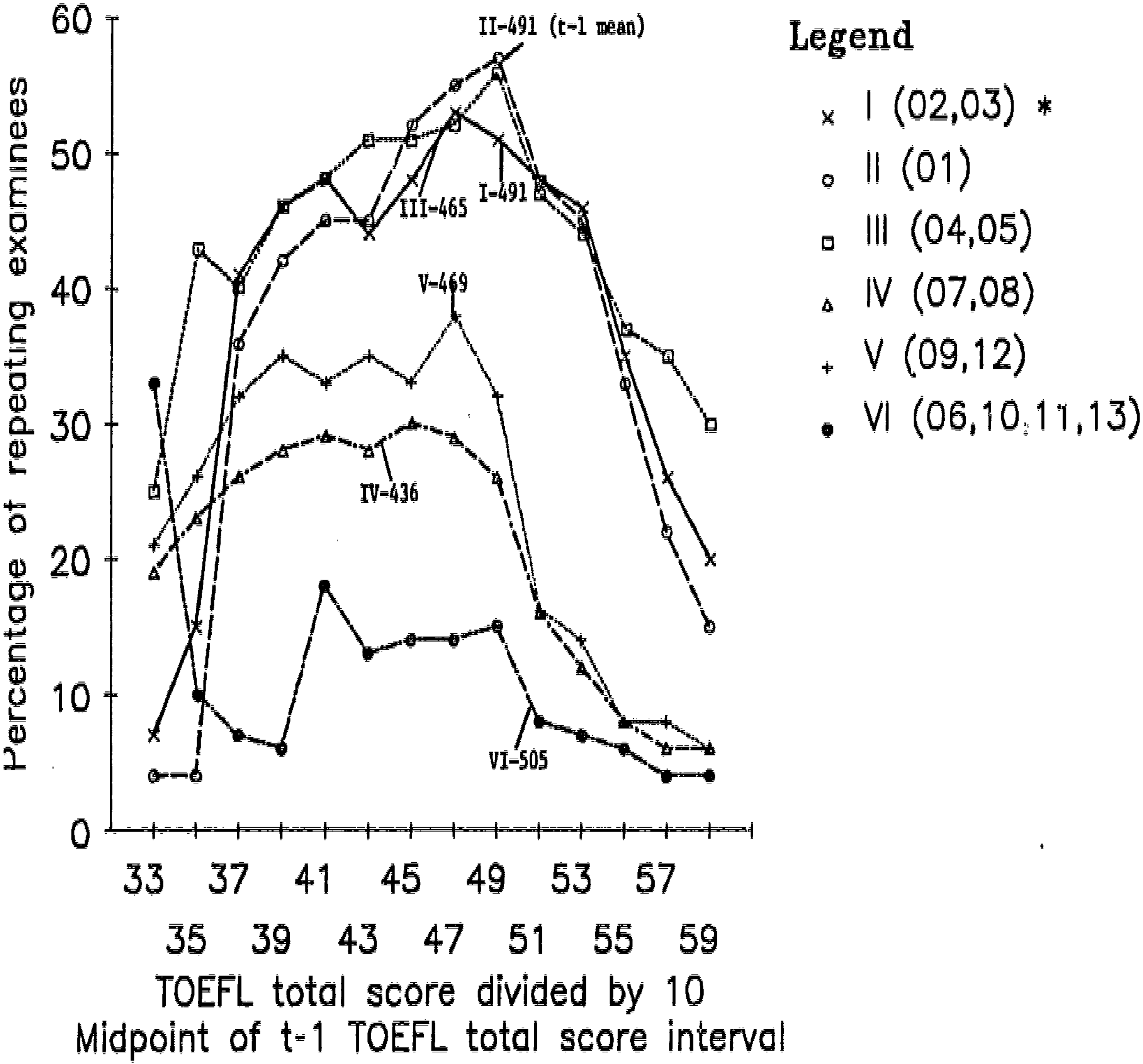
TOEFL Total	Consolidated analysis group*						
at initial	I	II	III	IV	V	VI	Total
testing	%	%	%	%	%	%	%
580 +	19.9	15.0	30.3	6.0	5.9	3.5	9.6
560-579	25.6	21.5	34.9	6.0	7.7	4.5	13.6
540-559	35.0	33.0	36.6	7.7	8.4	5.9	19.8
520-539	46.1	45.4	43.7	11.6	14.4	7.3	27.9
500-519	48.3	48.4	46.8	16.0	16.2	8.4	31.1
480-499	51.1	57.4	56.4	25.5	32.5	14.9	39.5
460-479	53.2	55.3	52.3	28.7	38.1	14.4	40.4
440-459	48.2	51.5	51.3	30.3	32.7	14.0	38.0
420-439	44.4	44.7	51.1	28.5	34.7	12.8	35.6
400-419	48.3	45.0	47.5	29.2	33.3	18.2	35.7
380-399	46.5	42.0	45.7	28.3	34.9	5.9	33.0
360-379	40.6	35.6	40.0	26.4	32.0	7.1	30.1
340-359	15.2	4.5	43.4	23.0	26.0	9.5	25.4
< 340	6.8	**	25.0	19.1	21.0	33.3	20.9
Total	49.7	44.5	47.1	23.5	22.1	8.6	30.0

Note: Underscored entries indicate percentages of repeaters in the score intervals that include the mean t-1 score for repeaters in the designated groups.

*I (Hong Kong and Korea, original 02 and 03), II (Taiwan, 01), III (Thailand and Japan, 04 and 05), IV (Middle East, 07 and 08), V (South America and Greece and Turkey, 09 and 12), VI (India, Nigeria and Ghana, Europe-Germanic, Europe-Romance, 06, 13, 10, and 11).

**Less than five cases.

Figure 5. Percentage of repeaters by initial score level on TOEFL for conso..dated analysis groups (20 percent sample)



* Entries in ( ) are original analysis groups (see Table 16).

group VI (European, Indian, African) examinees, regardless of their initial score levels, repeated TOEFL. Moreove, most of them took only one additional test, while the Asian contingents included many multiple repeaters.

Some Unresolved Questions

What accounts for the differences among analysis groups in incidence of repeated test-taking by initial score levels on TOEFL? Why should only about 10 percent of group VI examinees with initial scores slightly above 500 repeat TOEFL as compared to about 50 percent of Asian examinees at that initial score level? Do these differences in patterns of test taking behavior reflect differences in "perceived proficiency" for examinees at the same score level who have different linguistic-cultural backgrounds? Do U.S.-bound Asian students, as compared to, say, their European counterparts, perceive level of score on TOEFL to be more critical to the realization of their plans to study in the United States?

To what extent are differences in level of performance on TOEFL associated with differences in the realization of reported plans to study in the United States or Canada? Are there differences in realization of plans for non-repeating and repeating examinees? Research is needed to answer questions such as these--questions that relate to the dynamics of test repetiton in different subgroups within the population of International and Special Center examinees. Research is also needed to provide empirical evidence regarding the relationship between score-levels on TOEFL and measures of the functional communicative competence, general and academic, of ESL students (faculty ratings and/or self-ratings, for example).

What accounts for the apparent decrease in tendency to repeat among the lowest scoring first-time International and Special Center test takers? One plausible hypothesis is that many of these very low-scoring individuals who apparently did not repeat the TOEFL in an I&SC administration (a) may have enrolled in special ESL programs in order to improve their English proficiency, and (b) may have met institutional requirements through sucessful completion of those programs rather than by formal I&SC testing.

It is also possible that many of the low-proficiency I&SC nonrepeaters may have repeated the TOEFL again (unofficially) in test administrations supervised by individual institutions as part of the TOEFL Institutional Testing Program (ETS, 1983). There undoubtedly is some, perhaps considerable, overlap between the Institutional (INST) and I&SC examinee populations. At present there is no basis for estimating the extent of movement of examinees between the two testing program populations. However, overlapping (or movement between populations) may be due to a variety of logical patterns of test-taking behavior.

Movement between the two populations is probably most characteristic of first-time TOEFL takers (in I&SC or institutional test administrations) with low scores. For example, low-scoring first-time I&SC examinees may enroll directly, without further I&SC testing, for ESL instruction and take TOEFL a second time (or more) as part of an institutional test administration, with no further need to take TOEFL in an official I&SC test administration. Such a model would help to account for the finding that substantial percent- ages of very low-scoring first-time I&SC test-takers (for example, below 400 on TOEFL total) did not repeat as I&SC examinees. On the other hand, foreign nationals with recognizedly limited English proficiency may enroll in special ESL courses, and take TOEFL for the first time in an institutional test administration. They may later take TOEFL again, but in an I&SC administration to support an application for admission.*

Research is needed to establish both the degree of overlap and the typical flow patterns between the INST and the I&SC examinee populations. In the meantime, preliminary evidence regarding patterns of short-term test-taking behavior and score change for examinees tested for the first time as part of an institutional testing program is provided in the final section of this report.

* The TOEFL background question regarding previous experience with TOEFL is open-ended with respect to the auspices of previous test administrations.

Section X. Related Findings: Short-term Score-Change for Institutionally Tested Repeaters

The preceding sections provided information regarding test-taking patterns and TOEFL score change for examinees in 13 analysis groups (based primarily on country of origin and native language) (a) who were tested for the first time in an International and Special Center (I&SC) test administration between July 1977 and June 1980, and (b) who had taken TOEFL again at least one time in an I&SC test administration as of June 1982.

About 28 percent of first-time I&SC examinees accumulated at least one additional test record within 24 to 59 months after initial testing (mean number of tests = 2.9). The average interval between the initial and the last test administration was approximately one year (mean = 11.9 months). Taking into account the initial (time-1 or t-1) total score, the variables contributing most to prediction of final total score (time-T or t-T) were (a) being in the U.S. at the time of testing, (b) time interval between t-1 and t-T, and (c) number of times tested. Age, sex, and educational level at time of initial testing also contributed to prediction after control for initial score.

The average amount of change in TOEFL total score, over t-1 to t-T intervals averaging about one year, was 32 scaled score points, but there were substantial differences among the analysis groups with respect to mean change (from 18 points to 54 points). Some groups had substantially higher t-T means than expected based on their t-1 scores and other variables found to be associated with change (time interval, number of times tested, location, and so on); for other analysis groups, the opposite was true.

Because of the length of the follow-up period, the observed average score change for I&SC repeaters between the initial and last observed test administration may be thought of as reflecting "long-term" change associated with (a) tenure as members of the I&SC examinee population for the majority of repeaters and (b) the entire range of conditions that normally prevail between test- ings for individuals who repeat TOEFL in I&SC administrations.

This section presents information regarding "short-term" TOEFL total-score change for examinees who are assumed to have been enrolled in programs of instruction in English as a second language between testings. Based on data in TOEFL files, the examinees involved (a) were tested for the first time by an institution as part of the TOEFL Institutional Testing Program between July 1984 and July 1985, (b) indicated that they were planning undergraduate-level or graduate-level study in the U.S., (c) were in one of the analysis groups defined for the study, and (d) were identified as repeat- ers within the study period—that is, they had accumulated at least one additional test record as of August 1985, in an INST test administration, within one to 12 months following initial testing.

Information is not available regarding the actual status of the institutionally tested repeaters, their experience between test administrations, conditions governing opportunities for INST examinees to take TOEFL, and so on. However, it is assumed that most INST repeaters ordinarily are enrolled for intensive instruction in English as a second language and that their test records are obtained as part of institutional testing programs designed (a) to assess need for special instruction in English, (b) to evaluate progress following some period of ESL instruction, and/or (c) to certify the "readiness" of the individuals to undertake academic instruction in English.

If these assumptions are valid, average changes in TOEFL performance for institutionally tested repeaters may be thought as reflecting primarily average gains, over specified periods of time, associated with participation in the typical range of programs of formal instruction in English as a second language that are offered by the institutions involved. In any event, it is reasonable to assume that "intervening conditions" are more comparable for individuals and for analysis groups in samples of INST "short-term" repeaters than in samples of "long-term" I&SC repeaters.

Description of the Institutional Data

A total of 10,055 individuals from the countries included in the 13 basic analysis groups defined for the study were identified as first-time TOEFL takers in an institutional test administration, during the period July 1984 through August 1985. First-time INST examinees, generally, and in the great majority of analysis groups, had lower TOEFL means than did their I&SC counterparts (Table 17). This is consistent with the assumption that need for special instruction in English as a second language is associated with membership in the population of institutionally tested examinees, whereas the I&SC examinee population includes individuals with high as well as low levels of proficiency in English. The few exceptions to this pattern in Table 17 may be due primarily to sampling considerations.

Some 10.4 percent of the INST examinees were identified as repeaters—that is, were tested again, one or more times, <u>within one to 12 months following the initial testing</u> (see Table 18).*

The overall incidence of repeated test taking was lower in these INST samples than in the I&SC samples (about 28 percent) and the incidence of multiple test taking (three or more tests) was lower as well (compare data in Table 18 with data in Tables 2 and 3). However, such differences are consistent with expectation.

o The I&SC data reflect test taking in regularly scheduled monthly test administrations; individuals tested for the first time during a three-year period had from 24 to 59 months in which to repeat the TOEFL in a regular-

*Three analysis groups were represented by fewer than 10 repeaters: Group 06 (India), Group 10 (Germany, France, Netherlands, Denmark, Norway, Sweden), and Group 13 (Ghana, Nigeria). Data for repeaters in these very small groups are not reported separately, but were included in analyses involving repeaters without regard to analysis group.

Table 17

Means of TOEFL Examinees Tested for the First Time by An Institution
As Compared to Means for First-Time Examinees Tested in an
International and Special Center Administration

Analysis group		(N)	Listening Comp	Structure & Written Exp	Reading Comp & Vocabulary	Total
01	I&SC	36877	50.3	49.2	50.8	505.4
	INST	403	49.5	47.7	48.2	484.8
02	I&SC	23203	50.5	49.5	51.1	504.1
	INST	175	51.8	48.5	49.7	500.3
03	I&SC	10171	49.2	49.6	52.0	502.5
	INST	898	45.3	45.3	46.0	455.3
04	I&SC	9463	47.9	45.3	45.9	463.3
	INST	450	49.5	47.4	47.4	481.0
05	I&SC	14215	48.9	47.1	47.5	476.4
	INST	1891	45.4	45.3	44.3	450.1
06	I&SC	19157	54.0	55.8	57.4	557.2
	INST	96	53.2	51.2	51.8	521.8
07	I&SC	20830	49.4	44.4	43.9	459.0
	INST	3165	47.1	42.5	41.2	436.8
08	I&SC	31594	49.1	43.8	42.8	452.5
	INST	705	48.5	44.8	43.6	456.3
09	I&SC	15018	53.0	47.7	51.7	507.5
	INST	1478	49.4	45.7	48.4	478.4
10	I&SC	6449	60.9	56.4	57.1	581.5
	INST	139	58.8	54.6	54.2	558.9
11	I&SC	6131	5.1	54.3	57.2	558.3
	INST	343	53.0	51.6	53.0	525.2
12	I&SC	746	53.4	49.3	49.4	506.7
	INST	236	49.7	46.5	45.4	472.1
13	I&SC	20886	49.1	53.4	53.4	519.6
	INST	76	52.6	52.2	50.5	517.8

Note: Groups are: 01 (Taiwan), 02 (Hong Kong),; 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India) 07 (Saudia Arabia, Kuwait, Libya, Lebanon, Syria, Jordan, Iraq), 08 (Iran), 09 (Mexico, Chile, Colombia, Peru, Venezuela), 10 (Germany, Netherlands, Norway, Sweden, Denmark), 11 France, Italy, Spain, Portugal), 12 Greece, Turkey), 13 (Ghana, Nigeria). The INST samples are from TOEFL Institutional Testing Progam files for July 1984 through August 1985; the I&SC samples are those from International and Special Center files who were tested for the first time between July 1977 and June 1980.

Table 18

First Time Institutional Test Takers, July 1984-August 1985, by Analysis Group and Number of Times Tested During the Study Period

		Number of times tested						
Group*	N	1	2	3	4	5	6	7
01	403	369	26	4	4			
02	175	162	9	2	2			
03	898	762	117	16	3			
04	450	401	38	9	2			
05	1891	1591	210	64	17	4	5	
06	96	89	6	1				
07	3165	2901	180	50	23	9	1	1
08	705	631	54	17	3			
09	1478	1358	88	30	1	-	1	
10	139	134	5					
11	343	321	19	3				
12	236	212	19	5				
13	76	76						
Tot	10055	9007	771	201	55	13	7	1
		Percentage distribution**						
01		91.6	6.5	1.0	1.0			
02		92.6	5.1	1.1	1.1			
03		84.9	13.0	1.8	0.3			
04		89.1	8.4	2.0	0.4			
05		84.1	11.1	3.4	0.9	0.2	0.3	
06		92.7	6.2	1.0				
07		91.7	5.7	1.6	0.7	0.3	*	*
08		89.5	7.7	2.4	0.4			
09		91.9	6.0	2.0	*		*	
10		96.4	3.6					
11		93.6	5.5	0.9				
12		89.8	8.1	2.1				
13		100.0						
Total		89.6	7.7	2.0	0.5	0.1	0.1	*

Note. The patterns of test taking reflected in these data for the institutionally-tested (INST) samples differ markedly from those reported for the basic first-time International and Special Center (I&SC) samples involved in this study. Contributing factors are differences in the time periods covered and differences in INST and I&SC patterns of test administration.

* 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 06 (India), 07 (Saudia Arabia, Kuwait, Lebanon, Libya, Jordan, Syria, Iraq), 08 (Iran), 09 (Mexico, Venezuela, Peru, Colombia, Chile), 10 (Germany, Netherlands, Norway, Denmark), 11 (France, Italy, Spain, Portugal), 12 (Greece, Turkey), 13 (Ghana, Nigeria).

** Row percentages should equal 100.0 within rounding limits.

ly scheduled I&SC test administration. It is assumed that the length of the follow-up period was sufficient to cover the "tenure" of first-time test takers as "members of the I&SC-examinee population."

o INST data, on the other hand, reflect test-taking behavior observed during only a 12-month period, within the framework of patterns of test administration that undoubtedly varied across institutions with respect to frequency, purpose, degree of institutional control over examinee participation, and so on. Moreover, individuals tested for the first time in an INST administration in, say, July 1984, had a maximum of 12 months in which to repeat the test, and an unknown number of ad hoc opportunities to do so; individuals tested for the first time during the latter part of the 12-month study period, on the other hand, may have had no opportunity to repeat.

Data in Table 19 point up expected differences between the INST and I&SC repeater samples in mean time interval between first (time-1 or t-1) and last (time-T or t-T) testing (mean = 4.2 months for INST and 11.9 months for I&SC) and mean number of times tested (mean = 2.4 for INST and 2.9 for I&SC).

In addition, Table 19 shows that the great majority (about 85 percent) of INST examinees were tested by institutions located in the U.S. while only about 32 percent of their I&SC counterparts were in the U.S. when tested--for 8 of the 10 INST analysis groups, 90 percent or more of the repeaters were in the U.S. at the time of testing.

Initial and Last Observed TOEFL Total Means, and Mean Change, for INST Repeaters by Analysis Group

Table 20 shows the initial (time-1 or t-1) total score means and the last observed (time-T or t-T) total score means for INST repeaters, the mean interval between t-1 and t-T, and mean change ([t-T] - [t-1]), for the 10 analysis groups represented by 10 or more INST repeaters. Parallel data provided for I&SC repeater samples in the corresponding analysis groups (from Table 15) provide perspective for assessment of the INST findings. INST analysis groups were ranked in terms of mean change (higher to lower) and I&SC analysis groups were similarly ranked. The two sets of ranks are shown in Table 20.

Strong INST vs I&SC repeater population differences in level of performance on TOEFL (consistent with differences in the performance of all first-time test takers shown earlier in Table 17) are clearly evident. In every analysis group the first-time TOEFL performance of INST repeaters was lower than that of I&SC repeaters. The t-1 mean for all INST repeaters (424.1) was some 50 points lower than that for the I&SC repeaters (474.6). Moreover, the t-T mean for all INST repeaters (460.2) was some 14 points lower than the t-1 mean for I&SC repeaters (474.6); this was true for six of the 10 analysis groups.

Mean change for INST repeaters over time intervals averaging 4.2 months was some 36 points. For the respective analysis groups, over intervals averaging from 2.4 to 4.7 months, mean change varied from about 22 points (04

Table 19

Means On Selected Personal and Testing-Related Variables for INST Repeaters, by Analysis Group, with Comparative Data for I&SC Repeaters

Group		N	Intrvl	Variables* Numtsts	Edlevl	Sex	Center	Age
INST	01	29	4.7	2.3	0.66	0.69	0.97	27.0
I&SC	01	3070	11.6	2.9	0.83	0.41	0.08	25.1
INST	02	11	3.6	2.5	0.45	0.36	0.92	23.6
I&SC	02	1781	13.7	3.2	0.26	0.34	0.06	19.3
INST	03	129	2.4	2.2	0.36	0.41	0.60	23.3
I&SC	03	819	11.7	2.8	0.76	0.26	0.26	25.3
INST	04	46	3.9	2.2	0.72	0.35	1.00	22.8
I&SC	04	733	12.9	3.0	0.85	0.44	0.32	23.8
INST	05	280	4.9	2.4	0.60	0.48	0.69	23.3
I&SC	05	1346	11.3	3.2	0.50	0.34	0.40	23.8
INST	07	227	4.3	2.5	0.58	0.07	0.96	23.0
I&SC	07	814	12.6	2.7	0.34	0.07	0.65	21.8
INST	08	65	4.3	2.2	0.63	0.37	1.00	22.1
I&SC	08	1500	9.9	2.7	0.31	0.21	0.79	21.3
INST	09	112	4.5	2.3	0.54	0.45	0.99	23.4
I&SC	09	566	10.1	2.5	0.60	0.26	0.69	23.9
INST	11	22	3.8	2.1	0.52	0.59	1.00	22.9
I&SC	11	101	10.6	2.2	0.78	0.24	0.22	24.0
INST	12	23	3.5	2.2	0.52	0.13	0.96	23.1
I&SC	12	307	10.2	2.6	0.56	0.16	0.34	21.4
INST	Total	954	4.2	2.4	0.56	0.36	0.85	23.2
I&SC	Total	11612	11.9	2.9	0.57	0.30	0.32	23.0

Note: Groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 Japan; 07 (Saudi Arabia, Kuwait, Jordan, Lebanon, Iraq, Syria); 08 Iran; 09 Chile, Colombia, Mexico, Peru, Venezuela; 11 France, Italy, Spain, Portugal; 12 (Greece, Turkey). No data are shown for 06 India, 10 (Germany, Netherlands, Denmark, Norway, Sweden), and 13 (Ghana, Nigeria), represented in the INST sample by fewer than 10 repeaters with observations on the study variables. Data for I&SC repeaters are from Table 10.

* **Intrvl** (interval) in months between t-1 and t-t; **Numtsts** (number of times tested; **Edlevl** (educational level--graduate = 1, undergraduate = 0); **Sex** (M = 0, F = 1); **Center** (last test administered in U.S. = 1, not in U.S. = 0); **Age** in years at time of initial testing (t-1).

Table 20

Initial (t-1) and Last Observed (t-T) TOEFL Total Means for Institutional and I&SC Repeaters, Mean Time Interval Between t-1 and t-T, and Mean Score Change, by Analysis Group

Analysis group*	N	Mean interval (months)	T-1 mean	t-T mean	Difference between means time-1 to time-T and rank of group	
					Diff	Rank**
INST 01	29	4.7	439.0	466.3	27.3	(8)
I&SC 01	3070	11.6	490.7	508.8	18.1	(10)
INST 02	11	3.6	468.2	498.0	29.8	(7)
I&SC 02	1781	13.7	489.5	519.0	29.5	(7)
INST 03	129	2.4	414.0	438.6	24.6	(9)
I&SC 03	815	11.7	495.5	521.1	25.6	(9)
INST 04	46	3.9	430.1	451.8	21.7	(10)
I&SC 04	733	12.9	454.4	485.1	30.7	(8)
INST 05	280	4.9	430.9	464.6	33.7	(6)
I&SC 05	1346	11.3	470.8	503.1	32.3	(6)
INST 07	227	4.3	406.3	442.8	36.5	(4)
I&SC 07	814	12.6	439.7	485.4	45.7	(3)
INST 08	65	4.3	416.8	470.2	53.4	(1)
I&SC 08	150	9.9	434.4	488.6	54.2	(1)
INST 09	112	4.5	434.0	485.3	51.3	(3)
I&SC 09	566	10.1	465.8	513.9	48.1	(3)
INST 11	22	3.8	482.8	516.6	33.8	(5)
I&SC 11	101	10.6	530.7	566.9	36.2	(5)
INST 12	23	3.5	406.5	458.5	52.0	(2)
I&SC 12	307	10.2	475.0	518.0	43.0	(4)
ALL INST	954	4.2	424.1	460.2	36.1	
ALL I&SC	11612	11.9	474.6	506.8	32.2	

* Summary statistics are not reported for groups with N less than 10. Analysis groups are 01 (Taiwan), 02 (Hong Kong), 03 (Korea), 04 (Thailand), 05 (Japan), 07 (Saudi Arabia, Iraq, Lebanon, Syria, Kuwait, Jordan), 08 (Iran), 09 (Mexico, Colombia, Chile, Peru, Venezuela), 12 (Greece, Turkey), 11 (France, Italy, Spain, Portugal). Data not shown separately for 06 India, 10 (Germany, Netherlands, Norway, Denmark, Sweden), and 13 (Ghana-Nigeria)--due to N less than 10--are included in the respective totals.

* *These are ranks of groups in terms of differences between the means shown. INST and I&SC analysis groups were ranked separately from highest to lowest in terms of these mean difference values. Thus, for example, among INST repeaters in the 10 analysis groups, those from Taiwan ranked eighth in terms of mean score-change while among I&SC repeaters, those from Taiwan ranked tenth.

Thailand) to over 50 points (08 Iran; 09 Mexico, Chile, Colombia, Peru, Venezuela, and 12 Greece, Turkey).

To the extent that, as assumed, these institutional repeaters were enrolled in ESL instruction between the t-1 and the t-T test administrations, these mean changes may be thought of as suggesting patterns of average gains to be expected over comparable periods of time, for similar members of these analysis groups who normally elect or are selected into programs of ESL instruction such as those represented in the sample.

It is noteworthy that the analysis groups that demonstrated more (less) mean change under the more controlled between-administration INST conditions tended to be those that demonstrated more (less) mean change under the varied conditions prevailing between t-1 and t-T for I&SC repeaters. The ranking of INST analysis groups corresponded quite closely to the ranking of I&SC analysis groups in terms of mean change (rho = .89). Such systematic covariation, especially in view of the relatively small size of most of the INST repeater-samples and clear population differences in TOEFL performance, suggests basic differences among analysis groups (associated with differences in linguistic-cultural background) in the rate of acquisition of proficiency in English as a second language.

Predicting t-T Total Score for INST Repeaters

Being in the U.S. (immersed in an English-speaking environment) when tested (1 vs 0), time interval in months between first and last tests, number of times tested, age, sex (F=1, M=0), and educational level (graduate = 1, vs undergraduate = 0) were treated as independent variables, and final (t-T) TOEFL total score as the dependent variable, in a regression analysis based on data for the 954 INST repeaters. Interval between tests and being in the U.S. were the primary nontest contributors to prediction of final TOEFL score. Standard partial regression weights from the INST analysis and those from a parallel I&SC regression analysis are shown below:

Variable	Standardized weight	
	INST	I&SC
Time-1 total score	.80	.77
Interval in months (Intrvl)	.13	.12
U.S. testing (1,0)	.10	.11
Number of tests taken (Numtsts)	.03	.09
Age	-.04	-.09
Sex (F=1, M=0)	-.00	-.05
Educational level (Edlevel)	.04	-.01
Simple correlation of t-1 score with t-T score	(.78)	(.72)
Multiple correlation	.80	.76
Standard error of estimate	35.00	35.30
Number of cases	954	11,612

In evaluating the diminished role of number of times tested as a predictor of final TOEFL score (or change) in the INST analysis as compared to the I&SC analysis, it should be kept in mind (a) that INST repeaters had fewer opportunities to repeat and (b) that the meaning of this variable may be somewhat different for the INST samples than for I&SC samples. It is assumed that in the I&SC context, number of testings is more amenable to control by individual examinees (reflecting, say, motivation, effort, resources). In the INST context, on the other hand, number of testings is assumed to be more under the control of the testing institutions.

Analysis Group Differences in Observed vs Expected t-T TOEFL Total Means

The regression analysis provided an equation for predicting t-T TOEFL total score for INST repeaters, generally, taking into account t-1 TOEFL score, interval between tests, number of times tested, and the other study variables. The total-sample regression equation was as follows:

$$\begin{aligned}\text{t-T TOEFL score} = {} & .890\ (\text{t-1 score}) + 2.357\ (\textbf{Numtsts}) + 2.981\ (\textbf{Intrvl}) \\ & + 15.604\ (\textbf{U.S.-testing}) - .535\ (\textbf{Age}) - .452\ (\textbf{Sex}) \\ & - 4.179\ (\textbf{Edlevel}) + 66.146.\end{aligned}$$

Following the same procedures that were employed in the analysis of I&SC repeaters (see Table 15 and related dicussion), the regression equation based on data for INST repeaters without regard to analysis group was used to compute estimated t-T TOEFL total means for each of the 10 analysis groups. Table 21 shows the difference between the observed and the expected means for each INST analysis group. Positive discrepancies indicate that a group's t-T mean was higher than expected while negative discrepancies indicate the opposite. Analysis groups are ranked according to the difference between observed and expected means—algebraically, high to low. Results of the comparable discrepancy analysis for I&SC repeaters and a set of ranks for the corresponding I&SC analysis groups are also shown in Table 21.

The ranking of INST analysis groups in terms of performance relative to expectancy corresponded closely to the ranking of the I&SC samples from the same analysis groups (rho = .888 for the two sets of ranks). This result is consistent with the finding, previously reported, of close correspondence between rankings of analysis groups in the INST and the I&SC samples in terms of average change in TOEFL total score between initial and last observed test administrations. It thus lends further support to the inference that the observed analysis-group differences reflect basic group differences in characteristic rate of acquisition of proficiency in English as a second language.

Recapitulation and Evaluation of Findings Regarding Institutional Repeaters

The findings reported in this section are based on data from TOEFL files for a sample of individuals in selected national-linguistic analysis groups

Table 21

Difference between Observed and Predicted t-T TOEFL Total Means
for INST and I&SC Analysis Groups

Observed mean minus expected mean and
corresponding rank of group*

Group**	INST repeater samples			I&SC repeater samples		
	N	Diff	Rank	N	Diff	Rank
12	23	+14.2	(1)	307	+10.4	(3)
08	65	+13.7	(2)	1500	+ 5.6	(4)
09	112	+13.2	(3)	566	+12.4	(2)
11	22	+ 3.0	(4)	101	+24.0	(1)
05	280	- 1.3	(5)	1346	- 1.1	(6.5)
02	11	- 1.5	(6)	1781	- 1.1	(6.5)
03	129	- 4.0	(7)	819	2.3	(5)
07	227	- 4.3	(8)	814	- 1.9	(8)
01	29	- 8.0	(9)	3070	- 3.9	(9)
04	46	-14.6	(10)	733	- 7.4	(10)

* Expected means for the INST and the I&SC analysis groups were based on the respective total-sample regression equations.

** 12 (Greece, Turkey), 08 (Iran), 09 (Mexico, Chile, Colombia, Peru, Venezuela), 11 (France, Italy, Spain, Portugal), 05 Japan, 02 Hong Kong, 03 Korea, 07 (Saudi Arabia, Jordan, Kuwait, Lebanon, Iraq, Syria), 01 Taiwan, 04 Thailand.

who took TOEFL initially in an institutional test administration between July 1984 and August 1985. About 10 percent of these individuals accumulated at least one additional test record during that period; of these repeaters, about 80 percent had only one additional record in the study file for the period under consideration. The mean number of tests per repeater was 2.4, and the average time interval between the initial test administration and the last observed record was 4.2 months. Some 85 percent of the repeaters were tested by institutions located in the United States.

For 954 INST repeaters with data on all study variables, the average amount of change in TOEFL otal score, over t-1 to t-T intervals averaging slightly more than four months, was 36 scaled-score points. Time interval between t-1 and t-T, and being enrolled in a U.S. rather than a non-U.S.institution when tested were the nontest variables contributing most to prediction of t-T TOEFL total score (and change, t-1 to t-2, by inference) for the INST repeaters.

Mean change in TOEFL total was greater for some analysis groups than for others; groups registering higher average gains in score tended to have higher t-T means than expected for the average INST repeater with similar time-1 scores and nontest characteristics (that is, t-1 to t-T time interval, number of times tested, U.S. vs other institution, age, sex, and educational level—graduate vs undergraduate).

Whether based on mean change in TOEFL total score or the mean difference between observed and expected t-T scores, the ranks of national-linguistic groups among INST repeaters were found to correspond closely to the ranks for the same national-linguistic groups among International and Special Center (I&SC) repeaters. These systematic relationships obtained despite the fact that (a) most of the INST analysis groups were quite small (with potential for sampling error), (b) the circumstances associated with test repetition are quite different in the two test populations, and (c) clear population differences in initial level of proficiency (as measured by TOEFL). The consistency of findings thus strengthens the inference of differences among national-linguistic groups in characteristic rate of acquisition of proficiency in English as a second language.

It is assumed that the institutional repeaters were taking special instruction in English as a second language between test administrations. To the extent that this assumption is valid, the pattern of differences in average gains in TOEFL performance for INST analysis groups in these samples may be thought of as indicative of the pattern of differences that might be expected in similar groups of examinees who may participate in "typical" programs of special ESL instruction, under "typical" program conditions, over time intervals averaging about four months. Of course, the fact that most of the analysis groups were relatively small limits the accuracy of inferences from these data regarding the specific "average gains" to be expected for various groups.

Research is needed to obtain more detailed information regarding the actual circumstances of individuals who repeat the TOEFL in institutional test administrations, the characteristics of instructional programs involved, the conditions of test administration, and so on. Assuming the availability of such information, test-retest data from TOEFL Institutional Program files could be used to provide evidence regarding typical patterns of TOEFL score change for various national-linguistic groups in institutional settings classified according to, say, typical duration of ESL program, patterns of instruction, emphasis on English for general purposes as opposed to English for specific academic or occupatonal purposes, and so on.

References

Alderman, D. L. (1981). **Measurement error and SAT score change** (College Board Report No. 81-9). New York: College Entrance Examination Board.

Educational Testing Service (1983). **TOEFL test and score manual.** Princeton, NJ: Author.

Lord, F. M. (1967). "Elementary models for measuring change," in Harris, C. W., Ed. **Problems in measuring change.** Milwaukee, WI: The University of Wisconsin Press, pp. 21-38.

Rock, D. R. and Werts, C. (1979). **An analysis of time-related increments and/or decrements for GRE repeaters across ability and sex groups** (GREB No. 77-9R). Princeton, NJ: Educational Testing Service.

Swinton, Spencer S. (1983). **A manual for assessing instructional growth in language settings** (TOEFL Research Report No. 14). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1984). **Foreign nationals taking the GRE General Test during 1981-82: Selected Characteristics and Test Performance** (GRE Board Professional Report No. 81-23bP). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1983). **A review of research on the prediction of academic performance after the freshman year** (College Board Report No. 83-2 & ETS RR-83-11). New York: College Entrance Examination Board.

Wilson, K. M. (1982). **A comparative analysis of TOEFL examinee characteristics, 1977-79** (TOEFL Research Report No. 11). Princeton, NJ: Educational Testing Service.

TOEFL Research Reports currently available...

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L. D. Clark. November 1977.

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W. Pike. June 1979.

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.* Paul J. Angelis, Spencer S. Swinton, and William R. Cowell. October 1979.

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D. Clark and Spencer S. Swinton. October 1979.

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* Donald E. Powers. December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Spencer S. Swinton and Donald E. Powers. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.* John L.D. Clark and Spencer S. Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance.* Gordon A. Hale, Paul J. Angelis, and Lawrence A. Thibodeau. December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language.* Donald L. Alderman and Paul W. Holland. August 1981.

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L. Alderman. November 1981.

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* Kenneth M. Wilson. July 1982.

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL.* Kenneth M. Wilson. July 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting.* Donald E. Powers and Charles W. Stansfield. January 1983.

Report 14. *A Manual for Assessing Language Growth in Instructional Settings.* Spencer S. Swinton. February 1983.

Report 15. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students.* Brent Bridgeman and Sybil Carlson. September 1983.

Report 16. *Summaries of Sudies Involving the Test of English as a Foreign Language, 1963-1982.* Gordon A. Hale, Charles W. Stansfield, and Richard P. Duran. February 1984.

Report 17. *TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper.* Richard P. Duran, Michael Canale, Joyce Penfield, Charles W. Stansfield, and Judith E. Liskin-Gasparro. February 1985.

Report 18. *A Preliminary Study of Raters for the Test of Spoken English.* Isaac I. Bejar. February 1985.

Report 19. *Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English.* Sybil B. Carlson, Brent Bridgeman, Roberta Camp, and Janet Waanders. August 1985.

Report 20. *A Survey of Academic Demands Related to Listening Skills.* Donald E. Powers. December 1985.

Report 21. *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference.* Charles W. Stansfield. May 1986.

Report 22. *Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.* Kenneth M. Wilson. January 1987.