RESEARCH
REPORT

July 2001
RR-01-15

Overestimation of LPI Ratings for Native-Korean Speakers in the TOEIC Testing Context: Search for Explanation

Kenneth M. Wilson

Educational Testing Service

Statistics & Research Division
Princeton, NJ 08541

Overestimation of LPI Ratings for Native-Korean Speakers in the TOEIC Testing Context: Search for Explanation

Kenneth M. Wilson

With the collaboration of:
Steven A. Stupak
Korea International
Human Resources Development Center

July 2001

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 10-R
Educational Testing Service
Princeton, NJ 08541

ABSTRACT

Regression-based guidelines have been developed for predicting ratings of speaking proficiency in English as a foreign language (EFL), as assessed using the Language Proficiency Interview (LPI) procedure, from scores on the TOEIC® (Test of English for International Communication) Test. These guidelines, which reflect the regression of LPI rating on TOEIC scores in a combined sample composed of native speakers of Japanese, French, Spanish, and Arabic, respectively, were found to overestimate LPI rating when applied to data for native speakers of Korean. It was hypothesized that this outcome would be expected if development of EFL speaking proficiency (as assessed by the LPI procedure) tends to lag *relatively* more behind development of the proficiencies that are assessed by the TOEIC Test, for native-Korean-speaking EFL learners than for demographically comparable EFL learners in other national/linguistic settings (at least those represented in the study sample). An indirect assessment of this hypothesis was undertaken in a series of exploratory analyses involving section- and total-score means for native-language groups on the familiar, three-part version of the Test of English as a Foreign Language™ (TOEFL®). Findings were interpreted as being generally, albeit indirectly, supportive of the developmental lag hypothesis, as well as of working assumptions underlying introduction and analysis of TOEFL native-language reference group means as group analytic variables. Needed lines of inquiry are suggested.

Key words and phrases: TOEIC Test, LPI rating, target-language-learning rate, multilevel analysis, group analytic variables, TOEFL reference group means

ACKNOWLEDGEMENTS

The data for native-Korean speakers that were used in this study were provided by Steven Stupak, who conducted the LPIs during an extended stay in Korea. Essential indirect support for the study was provided by the ETS Research Division. Very useful reviews of earlier versions of this manuscript were provided by Brent Bridgeman, Hunter Breland, and Lawrence Stricker. These contributions are gratefully acknowledged.

TABLE OF CONTENTS

Page

Introduction 1
TOEIC Test Validity Studies 1
The Present Study 3
General Overview of the Present Study 3
Relevance of TOEFL Means for Hypothesis Testing 5
Working Assumptions, Empirical Evaluations, Interpretive Inferences 6
TOEFL Means as Group Analytic Variables 7
A brief methodological note 7
The present multilevel application 8
Findings 9
Phase I Findings 9
Analysis of Overprediction in the Korean Sample 12
Phase II Findings 14
Separate Analysis of TOEFL Section Means for Native-Language Reference Groups 14
Contribution of TOEFL Means to Prediction of LPI Rating in Multilevel Analyses 18
Analytic Rationale 18
Procedure 19
Simulated cross-validation 20
Findings of the Regression and Residual Analyses: Total Sample 21
Cross-Validation Findings 25
Discussion 27
Indirect Support for the Developmental Lag Hypothesis 28
Suggestions for Future Research 29
Research on EFL Learning Conditions 30
Research on Comparative EFL Learning Difficulty for Native Speakers of Other Languages 31
Suggested Lines of Inquiry 33
Models for National Surveys 35
References 39
Appendices 43
Notes 56

LIST OF TABLES

Page

Table 1 Descriptive Statistics for Study Variables in Designated TOEIC/LPI Samples 10

Table 2 Intercorrelations of Variables in the Korean Sample (above the diagonal) and Previously Studied Sample (below the diagonal), with Corresponding Descriptive Statistics 10

Table 3 Correlation Patterns for the Combined Study Sample (N = 738): Korea (N = 149) plus Total Previously Studied (N = 589) 11

Table 4 One-way Analysis of Variance Results: Mean Residuals for Designated Samples 13

Table 5 Mean Oralcy Indices and Corresponding TOEFL Means for TOEFL Examinees Classified by Native-Language Group (1990); 110 Language Groups Involved in this Analysis, Listed in Descending Order with Respect to Oralcy 15

Table 6 Mean Oralcy Scores and Corresponding TOEFL Score Means for Language Groups Represented in the TOEIC/LPI Sample, in Descending Order with Respect to Oralcy 18

Table 7 Intercorrelations and Descriptive Statistics for TOEIC Test Scores, TOEFL-Generated Group-Level Scores, and LPI Rating: Combined Sample (N = 738) 22

Table 8 Selected Results of Exploratory Regression and Residual Analysis Designed to Assess the Contribution of the Several Group-Level Measures to Prediction of LPI Rating in the Combined Sample (N = 738) 24

Table 9 Mean Residuals (LPI Rating minus Predicted Rating) when Prediction is Based on TOEIC Total Only, and on Designated TOEIC Score(s) Augmented by TOEFL-Generated Measures for Group-Level Listening Comprehension (LCt) and Knowledge of Structure and Written Expression (SWEt), Respectively 25

Table 10 Mean LPI Residuals by Language Group for Designated Half-Samples A and B, Respectively, when Regression Equations Developed in A Were Used to Predict LPI Values in B (A→B), and vice versa (B→A), with Associated ANOVA Results 27

Table 11 Classification of 40 Languages According to Empirically Determined Learning Difficulty for Linguistically Mature Native-English Speakers (U.S. Military Personnel) 32

Table 12 Expectancy of Meeting Graduation Standard in Designated Categories of Foreign Languages, by Designated Score Levels on the Defense Language Aptitude Battery (DLAB) 33

LIST OF TABLES

Page

Table 13 TOEFL Total Means for 1977, 1982, 1990, and 1996, respectively, for Native Language Reference Groups Selected to Represent Groups of Languages Identified by DLIFLC as Being *Most Difficult* and *Easiest*, respectively, for Native-English Speaking Language Students to Learn 34

Table A-1 TOEFL Total Means by Native-Language Reference Groups, Represented by $N > 24$ in Designated Testing Periods Between 1971 and 1996, Inclusive: Groups Listed in Descending Order by 1971 Mean 49

Table A-2 Descriptive Statistics and Stability Coefficients for TOEFL Means of 68 Native-Language Groups for Designated Testing Periods Between 1977 and 1996 51

Table A-3 Trends in Correlation Between Time-1 and Time-T Total Scores, by Number of Times Tested, for Repeaters Without Regard to Analysis Group, with Corresponding Descriptive Statistics 52

Table A-4 Stability Coefficients and Descriptive Statistics for Oralcy Indices circa 1977, 1982b, 1990, and 1996, Respectively, for 83 TOEFL Native-Language Contingents Represented in All Four Time Periods 53

Table B-1 Interrcorrelations and Descriptive Statistics for Two Half-Samples, A ($N = 368$) and B ($N = 370$) 55

INTRODUCTION

The TOEIC® (Test of English for International Communication) Test is a test of proficiency in English as a second language (ESL) administered by The Chauncey Group International™, Inc., (CGI) (e.g., CGI, 1999), but originally administered by the Educational Testing Service® (ETS®). It was developed by ETS at the request of the Japanese Ministry of International Trade and Industry, and introduced in Japan in 1979 to meet the need for a test of ESL proficiency designed for use by corporations and other agencies concerned with selecting, placing, or training employees for ESL-essential or ESL-desirable assignments (e.g., ETS, 1985a, 1986). Each year, the TOEIC program develops several new forms of the test for use in scheduled, Secure Program administrations in Japan and Korea, only. These test forms are used subsequently in some 25 other countries, in on-site, Institutional Program administrations conducted by TOEIC representatives or by representatives of client organizations. Regardless of the program, however, scoring and reporting of test scores are the responsibility of either CGI (for the Secure Program) or the respective local TOEIC representative agencies.

The TOEIC Test provides separate scores for ESL Listening Comprehension (LC) and Reading Comprehension (RC), each reported on a standard scale ranging between 5 and 495, as well as a Total score (range = 10 – 990), which is the simple sum of the reported LC and RC scores. An independent review of the TOEIC Test (Perkins, 1987) characterizes the TOEIC Test's psychometric properties, in part, as follows:

> In sum, the TOEIC is a standardized, highly reliable and valid measure of English, specifically designed to assess real-life reading and listening skills of candidates who will use English in a work context . . . In addition to being an integrative test, the TOEIC also appears to tap communicative competence in that the items require the examinee to utilize his or her socio-linguistic and strategic competence. (p. 82)

For detail regarding the TOEIC Test see, for example, CGI (1996, 1999) and ETS (1986).

TOEIC Test Validity Studies

In the basic validity study (Woodford, 1982) conducted to coincide with introduction of the TOEIC Test in Japan it was found that in the sample of native-Japanese speakers involved, level of performance on the TOEIC Test was relatively closely related (r = .83) to level of ESL speaking proficiency as elicited and evaluated through the Language Proficiency Interview (LPI) procedure (also known as the ILR Oral Proficiency Interview). This procedure for directly assessing target

language speaking proficiency was developed by the Foreign Service Institute (FSI) of the U. S. Department of State, and subsequently adopted by the Interagency Language Roundtable (ILR). (For historical perspective regarding development of the LPI procedure, as well as conceptually comparable procedures for the direct assessment of other language macro-skills, see ETS, 1982a; Lowe, 1987; Lowe & Stansfield, 1988. See Wilson, 1989, pp. 6-10, for development of a detailed rationale for use of LPI performance as a *context independent* surrogate for real-life observations of speaking proficiency in a target language.)

For present purposes, apart from the fact that it is a direct, interactional procedure with significant face validity as a measure of an examinee's ability to "exchange meaning conversationally using English as a target language," the most distinctive feature of the FSI/ILR procedure is that the linguistic behavior elicited is rated according to inherently meaningful (that is, behaviorally defined) levels on a "quasi-absolute proficiency scale" (after Carroll, 1967). The scale for rating proficiency includes descriptions for six basic levels (Level 0, Level 1, . . . , Level 5) ranging from *no proficiency in the language* (Level 0) through *proficiency equivalent to that of an educated native speaker* (Level 5), and five in-between levels (0 plus, 1 plus, . . . , 4 plus), which for statistical purposes are coded by adding ".5" to the corresponding basic levels. Hence, LPI ratings span an 11-point numerical scale (0, 0.5, 1.0, 1.5, . . . , 4.5, 5.0).

Woodford's (1982) findings involving Japanese examinees were confirmed for other language groups by findings of a second validity study (Wilson, 1989) concerned with referencing TOEIC scores to the LPI scale. This study found, for example, that zero-order correlations between LPI rating and TOEIC scores centered around .7 within samples of native speakers of Japanese, French, Spanish, and Arabic, respectively, tested in Japan, France, Mexico, and Saudi Arabia, respectively. [1] It was noted (Wilson, 1989, p. 44) that in assessing consistency of TOEIC/LPI relationships it is important to examine consistency across samples with respect to level and pattern of concurrent correlations—that is, consistency in degree of rank-order agreement between the scores of individuals on two measures. In addition, attention was called to the importance of questions about the extent to which ESL users/learners (of the type likely to be taking a test such as the TOEIC Test) who present particular TOEIC scores tend to exhibit about the same average level of LPI performance without regard to national/linguistic origin.

Accordingly, regression equations developed in the combined sample were used (Wilson, 1989) to generate estimated LPI ratings for members of the respective linguistic sub-samples.

Comparison of the observed and estimated LPI means indicated relatively close agreement from sample to sample between observed average performance on the LPI and the average predicted from scores on the TOEIC Test using the general sample equation; means of residuals (discrepancies between observed and estimated values) for the respective language-group samples tended to be smallest when a total-sample equation involving only TOEIC LC was used (Wilson, 1989, p. 8).[2] The regression-based guidelines developed using data for the combined sample were subsequently applied (Wilson & Chavanich, 1989) to data for 196 cabinet attendant trainees employed by an international airline in Thailand (Thai Air), using data provided by Kasren Chavanich, senior English instructor.[3] Findings for the Thai Air sample indicated close agreement at the mean for distributions of observed and expected LPI ratings based on the previously developed guidelines.

THE PRESENT STUDY

Regression-based guidelines (for inferring probable level of LPI-rated speaking proficiency from TOEIC scores) based on data for samples of TOEIC-takers from a limited number of language groups are not necessarily generalizable to any or all other national/linguistic subpopulations of educated, adult ESL users/learners who are likely to be tested with the TOEIC Test. The present study was undertaken initially to extend assessment of TOEIC/LPI relationships across samples differing in native language by analyzing TOEIC scores and LPI ratings collected in the course of operational TOEIC assessments involving native-Korean-speaking EFL/ESL users/learners (N = 149) who were tested in their places of employment in Seoul. Based on unexpected results of the planned analysis of the data (Phase I of the study), the scope of inquiry was substantially expanded in an effort to help explain the unexpected findings. The nature and scope of the initially planned inquiry and the subsequent inquiry (Phase II) are explained in the general overview that follows.

General Overview of the Present Study

The study involved two related but clearly distinguishable phases. In Phase I, as planned, regression and residual analyses were conducted to evaluate, among other things, the extent to which previously developed regression-based guidelines would be applicable in the Korean sample. The most noteworthy finding for present purposes was that previously developed equations for estimating LPI performance from TOEIC scores tended to overestimate LPI performance for the

sample of native-Korean speakers by almost half a level on the ILR scale—inconsistent with findings for the other language groups. Both sampling- and method-related explanations (such as systematic negative selection on EFL speaking proficiency, systematic rater bias, and so on) for this finding were considered and tentatively ruled out, primarily because of the absence of such bias in previous ratings by the experienced, periodically recalibrated interviewer/rater who conducted and rated the interviews under consideration (Wilson, 1989).[4]

Having tentatively ruled out method- and sampling-related explanations, attention turned to the possibility that a more substantive explanatory rationale might be warranted. More specifically, it was reasoned for working purposes that the observed overprediction of LPI rating for native-Korean speakers might be expected if, under usual EFL learning conditions, acquisition of oral EFL communication skills (as assessed, for example, by the LPI procedure) tends to lag relatively more behind development of facility in other language macroskills[5] (such as those measured by TOEIC LC and TOEIC RC, for example) for Korean EFL users/learners than for demographically comparable EFL users/learners in other national/linguistic contexts (at least those represented in the TOEIC sample).[6]

This working hypothesis, referred to hereafter as the "differential developmental lag" hypothesis, could not be addressed using the data available. However, in Phase II of this study the differential developmental lag hypothesis was addressed, albeit indirectly, by introducing for analysis selected summary data from the familiar TOEFL® (Test of English as a Foreign Language®) testing context (e.g., ETS, 1997). Unless otherwise indicated, the foregoing and subsequent references to the TOEFL test are to the familiar, three-part, paper-and-pencil version of that test, which was adapted from the original, five-part version that is now in the process of being superseded for international testing purposes by a computer-adaptive version (see, for example, ETS, 1997). For present purposes, general familiarity with the traditional, three-part TOEFL, the TOEFL scale, and so on, is assumed; however, limited detail regarding TOEFL scores may be found in the designated endnote.[7] The three-part TOEFL has sections labeled, respectively, Listening Comprehension (LC), Structure and Written Expression (SWE), and Reading Comprehension and Vocabulary (RCV). The data introduced were mean total and section scores for TOEFL native-language reference groups (contingents of international students planning to study in the United States or Canada, classified by self-reported native language).

The rationale for introducing TOEFL language-group means for an indirect assessment of the differential developmental lag hypothesis is developed in considerable detail in this general overview, which also provides a general description of the analyses involving those means that were undertaken to evaluate that rationale empirically. In the absence of guidance from previous research involving comparable data and conceptualizations—no reports of similar studies were located in ETS research archives or elsewhere—the analyses undertaken in Phase II of the study were necessarily exploratory in nature.

Relevance of TOEFL Means for Hypothesis Testing

Summary statistics for examinee subgroups defined by self-reported native language (and native country, as well) have been reported periodically for more than a quarter of a century (e.g., ETS, 1973, 1978, 1982b, 1985b, 1990, 1991, 1996, 1997). The relevance of the section means for testing the differential developmental lag hypothesis was predicated on a series of interrelated working assumptions about the meaning of observed differences among native-language contingents in level and pattern of average performance on the TOEFL test. Each of the sets of assumptions was subjected to empirical evaluation. First, general assumptions about the nature and potential usefulness of the information being provided by TOEFL means were evaluated, as were assumptions about their relevance for testing the differential developmental lag hypothesis. The foregoing empirical evaluations focused exclusively on means for a representative array of TOEFL native-language groups as reported for an arbitrarily selected testing period as the primary units of analysis. Results of these empirical analyses involving group-level TOEFL data only were interpreted as being generally supportive of both the corresponding networks of assumptions and the differential developmental lag hypothesis. Based on the generally positive results of the analyses involving only group-level data from the TOEFL testing context, selected TOEFL native-language group means (those for the six native-language groups represented in the TOEIC sample) were included (along with the TOEIC scores of the individuals involved) as "group analytic variables" in a series of exploratory multilevel analyses (e.g., Pedzahur, 1997, p. 688, to be considered in greater detail later). The assumptions, related lines of reasoning, and related empirical findings are provided below. A detailed presentation of the findings alluded to in the overview is provided in the "Findings" section of this paper.

Working Assumptions, Empirical Evaluations, Interpretive Inferences

First, it was assumed that differences among TOEFL native-language reference groups in average standing on the respective sections tend to reflect to some meaningful extent relatively stable differences among such contingents with respect to characteristic rates of acquisition of the corresponding aspects of proficiency under very generally comparable foreign-language learning conditions.[8] One necessary condition for any further consideration of these means as a source of valid information about the posited group-related differences in characteristic EFL learning rates is that they should be stable over time. Results of comprehensive, ancillary analyses reported in detail in Appendix A indicate clearly that the relative standing of language groups on the TOEFL Test, generally and by section, has remained relatively stable over at least the past 25 years. Such stability suggests that the observed means tend to be resistant to possible differential change across diverse national/linguistic EFL learning contexts with respect to selection-related and/or instruction-related variables.

Second, because the development of speaking proficiency in a target language presupposes development of the functional ability to comprehend utterances in that language, it was assumed for working purposes that higher (or lower) means for TOEFL LC relative to means on RCV and SWE (that is, sections not involving spoken stimuli) tend to reflect, to some meaningful extent, correspondingly higher (or lower) relative latent levels of EFL speaking proficiency. It appeared to follow logically that one necessary condition for accepting the differential developmental lag hypothesis would be met if the LC mean for native speakers of Korean tends to be lower than that of other native-language contingents, at least those represented in the TOEIC sample, with comparable scores on the nonlistening (SWE and RCV) sections. To evaluate this line of reasoning, TOEFL LC, SWE, and RCV means for 110 native-language contingents (tested in 1990–91, an arbitrarily selected time period) were employed as units of analysis. The LC mean was regressed on those for SWE and RCV, and the corresponding residual (observed LC mean minus predicted LC mean from the regression-weighted composite of SWE and RCV)—called an *Oralcy index*—was computed in an analysis to be reported in detail later. For present purposes, it is sufficient to note that the observed LC mean for native-Korean speakers was lower, relative to expectancy for contingents with comparable means for SWE and RCV, than that of any of the other native-language contingents, including contingents from the language groups represented in the TOEIC sample. Based on the results of ancillary analyses reported in detail in Appendix A, the rank-

ordering of groups in terms of TOEFL section means and the derived Oralcy index, respectively, has been relatively stable over the past quarter of a century.

Based on its consistency with the assumptions and lines of reasoning underlying the exploratory analysis, this finding was interpreted as tending to parallel conceptually the Phase I finding that native-Korean speakers in the TOEIC study sample tended to have lower observed LPI ratings, on average, than predicted from TOEIC scores using regression equations developed from data for a general, linguistically heterogeneous sample that did not include native speakers of Korean. These outcomes appeared to lend support to an assumption that inferences about differences in EFL learning patterns by native language drawn from observed differences in TOEFL means for the corresponding native-language reference groups may tend to be generalizable. More specifically, such inferences may tend to be generalizable to the corresponding national/linguistic populations of EFL users/learners, hence may also be generalizable to national/linguistic subpopulations of EFL users/learners who tend to take the TOEIC Test .[9] Such an assumption is implicit in the introduction of TOEFL native-language reference group means as a potential basis for explaining native-language-related outcomes involving measures on individuals in the TOEIC testing context.

TOEFL Means as Group Analytic Variables

In any event, the findings that have been reviewed thus far suggested that it would be useful to conduct a series of multilevel analyses involving data for the study sample. In these multilevel analyses, TOEFL means (for the six language groups represented in the study sample) were treated as independent, *group analytic* variables (measures of group-level patterns of performance, with zero within-group variance) along with the TOEIC scores of individuals in the study sample, in regressions involving LPI rating as the dependent or criterion variable.

A brief methodological note. A group analytic variable is a measure or index with ". . . analytic properties based on the aggregation of data collected on members of [given] groups (e.g., mean intelligence, motivation, anxiety)" (Pedzahur, 1997, p. 688). More generally, Pedzahur (1997, chap. 16) provides a thorough discussion of problems and issues involved in multilevel analyses involving both group-level and individual-level measures, as well as numerous references to multilevel studies. Multilevel analyses typically are designed to assess analytically (e.g., by means of multiple regression analysis) the net effect of group analytic variables (e.g., means for a

particular group of which an individual is a member) after having controlled for effects of the same (or similar) variables on the individual level. As Pedzahur (1997) notes:

> For example, in research aimed at studying the contextual effect of socioeconomic status (SES) of region of residence on voting behavior, each individual has two scores: one's own SES score and the mean SES of the region in which one resides. Voting behavior is regressed on both the individuals' SES scores and the SES means for the regions. The partial regression coefficient for the vector of SES means is taken as the contextual effect of the regions' SES. Similarly, in a study of achievement one may use individuals' mental ability scores as well as the mean ability of their class (school, school district). Again, the partial regression coefficient for the mental ability means is taken as the contextual effect of the groups' mental abilities on achievement. (p. 688)

Pedzahur provides extensive references to related research, much of which involved learning/achievement measures on individual students as dependent or criterion variables.

The present multilevel application. The multilevel analyses that were conducted embodied the general conceptual and methodological rationale described by Pedzahur (1997). For example, in the present study, each individual's set of TOEIC scores was supplemented by the set of TOEFL-generated means for his or her own native-language group. Thus, TOEFL LC, SWE, and RCV means for native-Japanese speakers were added to the record of each native-Japanese speaker in the TOEIC sample, those for native-Arabic speakers were similarly added to records for representatives of that language group in the study sample, and so on. Accordingly, each member of the study sample had two sets of scores, one consisting of his or her own scores on the TOEIC Test and the other consisting of scores on the several group analytic variables (reported TOEFL means for his or her particular language group). LPI rating was regressed on various combinations of TOEIC scores and the TOEFL-generated group analytic variables.

For purposes of this overview, it is sufficient to note that when combined-sample regression equations that included the group analytic variables with TOEIC scores were used to estimate LPI rating, rather than equations involving only TOEIC scores, results indicated that the group measures contributed incrementally to prediction of LPI rating in the combined sample. This was indexed by statistically significant partial regression weights, increases in multiple correlation, and improved fit at the mean between observed and estimated LPI distributions for members of the several language groups. The latter was especially noticeable for the sample of native speakers of Korean for whom the discrepancy between observed and predicted LPI rating was negligible when group measures

were introduced. This was found to be true as well under conditions of double cross-validation involving arbitrarily defined half samples.

FINDINGS

The organization of findings generally follows the developmental pattern outlined above. Additional detail regarding procedures is provided as needed in connection with the presentation of detailed findings for the Phase I and Phase II analyses that have been generally described in the preceding section.

Phase I Findings

Detailed results of descriptive and residual analyses involved in Phase I of the study are provided in this section. Table 1 summarizes results of a preliminary descriptive analysis of data collected from the language groups represented in the sample. The unexpected finding, alluded to in the overview above, is that although the TOEIC Total mean for the Korean sample (M=679) was higher than that of the total sample previously studied (M=615), the mean LPI rating for the Korean sample (M=1.6) was lower by almost a half-level on the LPI scale than that of the previous sample (M=2.0).

Intercorrelations of the variables were as shown in Table 2; means and standard deviations from Table 1, below, are included for perspective. Coefficients (and descriptive statistics) for the Korean sample are presented above the diagonal; coefficients and descriptive statistics shown below the diagonal are for the combined sample (N = 589) previously studied, made up of native speakers of French, Japanese, Arabic, Spanish, and Thai, respectively, assessed in France, Japan, Saudi Arabia, Mexico, and Thailand, respectively. For reasons noted earlier, it is noteworthy that the observed coefficient for TOEIC LC versus LPI rating (r = .70) is comparable to that observed for TOEIC Total score with LPI rating (r = .69). Table 3 shows similar statistics for the total sample (N = 738), which includes data for the native-Korean speakers.

Table 1

Descriptive Statistics for Study Variables in Designated TOEIC/LPI Samples

	LPI rating			TOEIC test score					
				LC		RC		Total	
Sample	N	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd
France[#]	56	2.3	.64	428	74	389	48	817	113
Japan[#]	285	1.9	.67	316	83	302	77	618	151
Mexico[#]	42	1.7	.62	262	106	237	104	499	204
Saudi[#]	10	2.0	.93	304	107	184	114	489	217
Thai[##]	196	2.1	.66	319	54	265	53	586	90
Total	589	2.0	.68	323	84	292	81	615	155
Korea	149	1.6	.51	346	54	333	64	679	102
Combined	738	1.9	.71	328	79	301	79	628	148

[#] Data for the basic TOEIC/LPI calibration study (Wilson, 1989; see pp. 30 ff. for detail regarding the Japanese sample, pp. 43 ff. for detail regarding the French, Mexican, and Saudi samples). The samples are designated here according to the country in which the assessment took place. The samples were composed exclusively of native speakers of the corresponding (dominant) languages, namely, French, Japanese, Spanish, Arabic, Thai, and Korean, respectively. These two related sample designations are used interchangeably hereafter. Generally speaking, in samples of TOEFL examinees, native language tends to be heavily nested in the native country (e.g., Wilson, 1982).

[##] Data roughly concurrent with data for the calibration study, received too late for inclusion in the basic TOEIC/LPI calibration (Wilson, 1989) but subsequently described and analyzed (Wilson & Chavanich, 1989).

Table 2

Intercorrelations of Variables in the Korean Sample (above the diagonal) and Previously Studied Sample (below the diagonal), with Corresponding Descriptive Statistics

	LPI	TOEIC Test score			Korea (N = 149)	
Variable	rating	LC	RC	Total	Mean	Sd
LPI rating	1.00	.48	.50	.57	1.6	.5
TOEIC LC	.70	1.00	.50	(.84)	346	54
TOEIC RC	.59	.76	1.00	(.89)	333	64
TOEIC Total	.69	(.94)	(.94)	1.00	679	102
Previously studied samples (N = 589)						
Mean	2.0	323	292	613		
SD	.7	84	81	155		

Note: The below-diagonal data are for a combined sample made up of previously studied samples of TOEIC examinees with LPI ratings and TOEIC scores resulting from operational assessments in Japan, France, Mexico, and Saudi Arabia (Wilson, 1989) and Thailand (Wilson & Chavanich, 1989). Coefficients in parentheses are spuriously high, part-whole coefficients.

Table 3

Correlation Patterns for the Combined Study Sample ($N = 738$): Korea ($N = 149$) plus Total Previously Studied ($N = 589$)

	TOEIC Test score				Descriptive statistics	
Variable	LC	RC	Total	LPI Rating	Mean	Sd
Listening	--	.73	(.93)	.63	328	79
Reading		--	(.93)	.51	301	79
Total			--	.61	629	148
LPI rating				--	1.9	.71

Note: The samples previously studied (combined $N = 589$) are TOEIC examinees with LPI ratings and TOEIC scores resulting from operational assessments in Japan, France, Mexico, and Saudi Arabia (Wilson, 1989) and Thailand (Wilson & Chavanich, 1989). Coefficients in parentheses are spuriously high, part-whole coefficients.

TOEIC/LPI correlations generally tended to be lower in the Korean sample than in the previously assessed total sample. The correlation between TOEIC LC and TOEIC RC was also lower in the Korean sample than in the previously studied sample; this was true as well for the combined study sample (which includes data for native-Korean speakers). Moreover, in the Korean sample, the observed coefficient for TOEIC RC with LPI rating is somewhat larger than that for TOEIC LC and inconsistent with the pattern for previously assessed samples, and the coefficient for TOEIC Total with LPI rating is larger than the corresponding coefficients for LC and RC, respectively. In evaluating these findings it is pertinent to recall that observed correlations between variables are sensitive to restriction-of-range effects. For example, correlations tend to be lower in samples with a restricted range of performance due to direct or indirect selection on the variables involved than in samples that are not similarly restricted. And, it is evident that TOEIC score dispersions for the Korean sample tend to be more restricted than are those for the other samples. For example, the standard deviation for TOEIC LC is 54 in the Korean sample as compared to 84 in the previously studied sample ($N = 589$), and the corresponding means are 346 and 323, respectively.

Analysis of Overprediction in the Korean Sample

Although average TOEIC scores for the Korean sample (e.g., TOEIC Total mean = 679) were found to be somewhat higher than the corresponding averages for the previously studied sample (TOEIC Total mean = 613), the average LPI rating for the Korean sample (1.6, or approximately Level 1+) was lower than that for the previously studied sample (2.0, or approximately Level 2). This suggests that previously developed regression equations may tend to overestimate LPI rating in samples of native-Korean-speaking TOEIC examinees. To provide a systematic assessment of the foregoing inference, regression equations for predicting LPI rating from TOEIC scores were applied to data for members of the combined study sample ($N = 738$). These were the equations developed in the TOEIC/LPI-calibration study involving data for the French, Japanese, Saudi, and Mexican samples (for detail, see Wilson, 1989, p. 71). The equations were used to predict LPI rating using the three possible TOEIC score options: LC only, LC and RC treated as separate predictors, and Total (the simple sum of LC and RC). Corresponding residual values (that is, LPI rating minus predicted LPI rating corresponding to each of the equations) were computed. One-way analysis of variance was then used to assess differences among groups with respect to mean residuals. The predictive equations involved and the selected results of the corresponding residual analysis are shown in Table 4. For these and subsequent analyses concerned with consistency of fit between distributions of observed and predicted LPI ratings, data for four different samples of native-Japanese speakers making up the total Japanese sample ($N = 285$) were analyzed separately.

For present purposes, it is sufficient to note that the most consistently large average discrepancy between LPI rating and the rating estimated from TOEIC score(s) is associated with the Korean sample with mean residuals of approximately -.4 for each equation. For reasons previously noted, it is of some interest that (judging from the size of the corresponding F values, for example) fit between distributions of observed and predicted ratings was slightly better for the equation involving TOEIC LC alone (LCres) than for the equations involving LC and RC, either as independent predictors (LRres) or as a simple sum in TOEIC Total (TOTres).

Table 4

One-way Analysis of Variance Results: Mean Residuals for Designated Samples

Sample		Mean of residuals		
	N	LCres	LCRCres	TOTres
France	56	-.17	-.16	-.15
JpIIST-84#	66	-.07	-.07	-.07
JpIIST-86#	55	-.06	-.06	-.07
JpIIST-87#	42	.07	.05	.03
JpTOEIC-85#	122	.04	.04	.03
Saudi	10	.14	.30	.44
Mexico	42	.14	.23	.18
Thai	196	.18	.23	.28
Korea	149	-.41	-.43	-.44
Total	738	-.04	-.02	-.02
F		19.6	25.3	29.7
P < .0001	(all)	(df 8,729)		
Eta^2		.177	.217	.246

Note: The three residual variables and the predictor(s) involved were as follows:

LCres	= LPI - LPI predicted from TOEIC LC only
LRres	= LPI - LPI predicted from the LC/RC equation
TOTres	= LPI - LPI predicted from TOEIC Total

The equations used were developed using data for the French, Japanese, Saudi, and Mexican samples. Accordingly, the results for the Thai as well as the Korean sample represent cross-validation applications of those equations. Negative residual values represent overprediction; positive residual values indicate underprediction. Thus, for example, the -.41 for LCres for the Korean sample indicates that when LPI rating was estimated from TOEIC LC only, the predicted values exceeded the observed values by the corresponding amount, on the average.

These four subsamples collectively constitute the Japanese sample ($N = 285$) for which means and standard deviations are reported in Table 1, above. They are identified separately in this analysis—and in some subsequent analyses—primarily to provide more comprehensive perspective on consistency of agreement between means for observed and predicted LPI ratings across samples within, as well as across, language groups.

Phase II Findings

As indicated in the general overview, Phase II findings are those resulting from (a) ad hoc regression and residual analyses involving only the means of native-language reference groups on the TOEFL and (b) multilevel regression and residual analyses designed to evaluate the contribution of TOEFL means treated as group analytic variables, along lines described in the preceding section. The presentation of findings (and related procedures, as needed) follows the foregoing pattern.

Separate Analysis of TOEFL Section Means for Native-Language Reference Groups

Based on assumptions and lines of reasoning developed in detail in the overview section above, ad hoc analyses focused exclusively on selected data from the TOEFL testing context were conducted in order to provide an indirect assessment of the differential developmental lag hypothesis. More specifically, the analyses were designed to assess the extent to which native-Korean speakers who take the TOEFL tend to earn relatively lower LC means (and, by inference, probably tend to have correspondingly lower average levels of EFL speaking proficiency) compared to members of other language groups, including those represented in the present study, after controlling for average performance on the other two TOEFL sections. To assess this hypothesis, TOEFL section means for 110 contingents of native-language examinees tested between 1989 and 1991 (hereafter, 1990) were selected as the primary units of analysis.

The TOEFL LC mean was treated as the dependent variable and the means for SWE and RCV were treated as the independent variables in a regression analysis. The resulting equation was used to compute a predicted LC value for each contingent. The corresponding residual value (LC mean minus predicted LC mean) was then computed. The distribution of residuals, with a mean of .00 (by definition) and a standard deviation of 2.4, was converted to a standard scale (mean = 50, standard deviation = 10). The resulting converted value was called an Oralcy index. Table 5 shows TOEFL section means and corresponding Oralcy indices for all groups involved in the analysis. It can be seen that Oralcy indices ranged between 29 and 74 (rounded). For present purposes, the most noteworthy aspect of this range is that it was anchored at the lower end by the contingent of native-Korean-speaking EFL learners in the TOEFL sample involved in the analysis. This finding indicates that among the 110 native-language contingents who took the TOEFL test in 1990, native-Korean speakers earned lower average scores on the measure of

listening comprehension (TOEFL LC) than did members of any other language group, after controlling for level of developed proficiency in reading comprehension and vocabulary, and knowledge of English language structure and written expression, as reflected in the corresponding TOEFL section means.

Table 5

Mean Oralcy Indices and Corresponding TOEFL Means for TOEFL Examinees Classified by Native-Language Group (1990); 110 Language Groups Involved in this Analysis, Listed in Descending Order with Respect to Oralcy

		TOEFL mean			
Language	Oralcy	LCt	SWEt	RCVt	TOTt
Swedish	73.9	62	58	57	591
Icelandic	73.9	60	55	54	563
Danish	72.7	62	59	57	593
Hebrew	72.4	59	54	53	551
Dutch	71.2	62	59	58	598
Norwegian	69.9	60	57	55	573
Finnish	65.5	60	58	57	584
Samoan	65.5	55	51	49	516
German	64.3	60	59	57	586
Yapese	63.8	56	52	52	531
Lao	63.6	53	47	48	491
Farsi/Persian	62.5	54	50	49	509
Serbo-Croatian	62.5	57	54	54	551
Palauan	62.4	52	47	46	482
Kurdish	61.1	53	49	48	499
Malaybahasa	61.0	56	53	53	540
Cambodian	60.9	52	47	47	487
Fijian	60.1	56	55	52	542
Armenian	59.9	55	53	51	530
English	59.8	59	58	58	584
Czech	59.7	57	55	55	558
Hungarian	59.7	57	55	55	558
Slovak	59.7	57	55	55	558
Guarani	59.2	53	48	50	500
Tagalog	58.3	58	57	57	573
Cebuano	58.3	58	57	57	571
Polish	58.3	56	54	54	547
Greek	57.3	54	53	50	526
Arabic	57.0	51	48	46	480
Romanian	56.9	57	56	56	564
Ponapean	55.8	53	52	49	511
Afrikaans	55.8	56	56	54	554
Pidgin	55.8	56	56	54	552

(*table continues*)

Table 5—*Continued*

Language	Oralcy	TOEFL mean LCt	SWEt	RCVt	TOTt
Russian	55.6	55	54	53	540
Maltese	54.4	60	62	61	607
Marshalles	54.2	48	45	42	451
Vietnamese	54.0	53	51	51	513
Bulgarian	52.8	55	55	54	548
Ilocano	52.6	54	53	53	530
Spanish	52.3	54	52	54	534
Tulu	51.3	57	58	58	577
Sinhalese	51.3	55	55	55	553
Tibetan	51.3	55	55	55	547
Indonesian	51.1	51	49	49	496
Siswati	50.5	56	59	55	568
Shona	50.4	57	60	57	583
Turkish	50.1	52	52	50	516
Tigrinya	50.1	52	52	50	512
Sindhi	50.0	55	56	55	554
Amharic	49.9	51	50	49	500
Gujarati	49.8	54	54	54	540
Basque/Euskara	49.7	55	55	56	553
Somali	49.7	50	48	48	484
Setswana	49.1	55	58	54	557
Urdu	48.5	52	52	51	518
Javanese	48.4	50	49	48	491
Italian	48.4	56	57	58	569
Sesotho	47.9	55	59	54	563
Zulu	47.6	55	58	55	560
Hindi	47.3	57	60	59	587
Catalan	46.9	55	56	57	561
Portuguese	46.9	53	53	54	534
Bemba	46.2	56	60	57	579
Trukese	45.8	46	45	42	443
Yoruba	45.8	54	56	55	548
Marathi	45.8	57	60	60	590
Other	45.7	52	53	52	522
Kusaiean	45.7	47	46	44	454
Burmese	45.5	51	51	51	513
French	45.5	54	55	56	551
Chichewa	44.8	55	59	56	567
Kikuyu	44.8	55	59	56	563
Kashmiri	44.7	56	60	58	582
Punjabi	44.4	53	55	54	540
Malayalam	44.3	56	59	59	579
Tamil	44.3	56	59	59	578
Ga	43.2	55	59	57	569

(*table continues*)

Table 5—*Continued*

		TOEFL mean			
Language	Oralcy	LCt	SWEt	RCVt	TOTt
Swahili	43.2	53	56	54	542
Tongan	55.3	50	46	46	475
Efik	43.0	52	54	53	529
Assamese	42.9	55	58	58	570
Ulithian	42.8	53	55	55	544
Chinese	42.8	51	52	52	509
Sudanese	42.7	49	49	49	490
Thai	42.7	49	49	49	489
Iboicbo	41.7	52	55	53	532
Wolof	41.6	50	52	50	506
Berber	41.5	49	50	49	492
Japanese	41.5	49	50	49	485
Nepali	41.4	52	54	54	535
Khalkha	41.3	50	51	51	505
Pushtu	41.3	50	51	51	504
Twifante	40.4	55	60	58	575
Hausa	40.3	51	54	52	526
Kannada	40.1	55	59	59	577
Lingala	40.0	48	49	48	483
Luo	39.1	53	58	55	555
Bengali	38.6	50	52	52	512
Telugu	38.5	53	56	57	553
Akan	37.6	53	58	56	557
Malinkemandingo	37.4	49	52	50	502
Galla	35.9	49	52	51	510
Malagasy	35.8	50	53	53	523
Unknown	35.7	48	50	50	493
Ewe	34.9	52	58	55	551
Mende	34.7	51	56	54	537
Fulani/Fulu	34.5	48	51	50	497
Lubalula	33.2	48	52	50	502
Oriya	31.5	52	57	58	556
Korean	28.6	48	52	53	505

Note: Oralcy is the residual variable LCt minus LCt predicted from SWEt and RCVt, standard scaled (M = 50, SD = 10), where LCt, SWEt, and RCVt, respectively, are means for the language groups on corresponding TOEFL sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary, respectively. The TOEFL means involved are as reported in ETS (1991, p. 23). To evaluate the stability of the Oralcy index over time, corresponding indices were computed for similarly defined native-language contingents represented in each of four testing periods (1977, 1982, 1990, and 1996, respectively). Results (reported in detail in Appendix A) indicated substantial overall stability in the relative standing of native-language contingents on this index, as well as in the means involved in its derivation. Oralcy indices based on the 1990 data and corresponding TOEFL means for the six language groups involved in the present study are summarized for convenience in Table 6.

Table 6

Mean Oralcy Scores and Corresponding TOEFL Score Means for Language Groups Represented in the TOEIC/LPI Sample, in Descending Order with Respect to Oralcy

		Group Analytic Variables				
		Oralcy	TOEFL mean (1990)			
Country	Language	Index	LCt	SWEt	RCt	TOTt
Saudi Arabia	Arabic	57	51	48	46	480
Mexico	Spanish	52	54	52	54	534
France	French	46	54	55	56	551
Thailand	Thai	43	49	49	49	489
Japan	Japanese	42	49	50	49	485
Korea	Korean	29	48	52	53	505

Note: LCt = mean TOEFL Listening Comprehension, SWEt = mean TOEFL Structure and Written Expression, RCt = mean TOEFL Reading Comprehension and Vocabulary, and TOTt = mean TOEFL Total.

Contribution of TOEFL Means to Prediction of LPI Rating in Multilevel Analyses

For each language group represented in the TOEIC sample, the corresponding Oralcy index and the several TOEFL means shown in Table 6 were assigned to individual members of the corresponding language groups in the TOEIC sample as supplemental (group-level) scores for the corresponding group analytic variables. For example, scores of 57 (Oralcy), 51 (mean TOEFL Listening Comprehension, or LCt), 48 (mean TOEFL Structure and Written Expression, or SWEt), 46 (mean TOEFL Reading Comprehension and Vocabulary, or RCt), and 480 (mean TOEFL Total or TOTt) were added to the record for each member of the native-Arabic-speaking Saudi Arabia sample. Similarly, scores of 52, 54, 52, 54, and 534 (see Table 6, above) were added to the record for each member of the native-Spanish-speaking sample assessed in Mexico, and so on.

Analytic Rationale

To assess how the means on the three TOEFL sections contribute to prediction of LPI rating in the combined sample, five sets of exploratory analyses were conducted, each involving three separate regression and residual analyses. These analyses were designed to permit effects associated with the various group analytic variables (native-language group TOEFL means) after controlling for the aspects of proficiency measured by individuals' scores on the TOEIC Test. As

noted by Pedzahur (1997, p. 688), regressing a criterion measure on sets of independent variables that include both group analytic measures (in the present instance, the reported TOEFL means for language groups) and individuals' scores on the same or related measures (in this instance, scores on the TOEIC Test) permits assessment of possible contextual effects associated with the group analytic variable(s) after controlling for effects associated with the scores of individuals on the same or similar measures under consideration. Such effects may be inferred from the presence of statistically significant partial regression coefficients for the group variables when included in sets with TOEIC scores, and larger multiple correlation coefficients for sets of predictors that include group measures, than for sets involving only individuals' TOEIC scores. If these conditions are found to obtain, it becomes meaningful to evaluate the net effect of the "net adjustments" specified by the relative weighting of the group analytic variables on the mean predicted criterion values for the respective groups. Such an evaluation might involve, for example, comparing the results of one-way analyses of variance for differences in means of groups with respect to distributions of residuals (LPI minus predicted LPI) corresponding to equations that include group analytic variables, on the one hand, with results corresponding to equations including only the TOEIC scores of individuals in the sample.

Procedure

All procedures were exploratory in nature, given the novel treatment of TOEFL means as group analytic variables—that is, variables thought of as reflecting characteristic differences associated with native language in the relative development of different aspects of EFL proficiency. The overall line of inquiry was designed to shed light on the contribution of the several group analytic variables to prediction of LPI rating when included with each of the three possible TOEIC score options, namely, Total only (LC and RC summed), LC only, and LC and RC as separate predictors. As an initial step, all three group-level scores (LCt, RCt, and SWEt), as a fixed set, were included, in turn, with each of the TOEIC score options noted above. Without regard to the TOEIC score(s) involved, beta (standard partial regression) weights for group listening comprehension and "knowledge of EFL structure" (LCt and SWEt), respectively, were found to be statistically significant. However, the weight for group-level reading comprehension (RCt) was found to be small and only marginally significant statistically. Based on these results, a decision was made to explore the contribution of both (a) an unweighted average of SWEt and RCt (the two nonlistening

group variables) labeled *Write*, and (b) SWEt only, when included with the respective TOEIC score options, namely, TOEIC LC only, TOEIC LC and RC, and TOEIC Total. For perspective in evaluating outcomes involving separate treatment of TOEFL section means as group measures, TOEFL Total mean (TOTt) was also introduced into a series of regressions with the three TOEIC score options as a group-level measure for general level of EFL proficiency. For each of the 15 resulting total-sample regression equations, the corresponding residual (LPI rating minus predicted LPI rating) was computed for individuals in the combined sample. Finally, one-way analysis of variance was used to permit a general evaluation of the extent to which differences among language groups on the respective residual variables tended to be more or less pronounced as a function of the particular combination of scores used to generate predicted LPI rating (especially TOEIC scores only versus TOEIC scores and scores for group variables).

Simulated cross-validation. There was marked improvement in fit between observed and predicted LPI distributions when group-level variables were included with TOEIC scores in the total-sample analysis. Since data were not available for another sample, the stability of the prediction systems involved could not be assessed directly. However, it was possible to examine the stability of prediction from equations that include the group measures by conducting simulated cross-validations involving selected combinations of TOEIC scores and group analytic measures using only the data at hand. Two validation samples (labeled "A" and "B") were defined by arbitrarily treating as different samples the first n/2 and the second n/2 cases, respectively, from each of the six language groups represented in the study sample as indicated below (where 1 represents the first n/2 cases, and 2 represents the second n/2 cases in the study file for each language group):

Half	French		Japanese		Arabic		Spanish		Thai		Korean	*N*
A =	1	+	2	+	1	+	2	+	1	+	2	= 368
B =	2	+	1	+	2	+	1	+	2	+	1	= 370

In each of the half-samples thus defined, LPI was regressed in turn on TOEIC LC, TOEIC LC and RC, and TOEIC Total. LPI was then regressed on the respective TOEIC options, with the two group analytic variables (LCt and SWEt) identified as making the primary incremental contribution to prediction in the total sample. Equations developed in A were used to generate predicted LPI values for members of B (A→B applications) and vice versa (B→A applications). In

each of the two half-samples, differences by language group with respect to the corresponding residuals were then subjected to one-way analysis of variance to evaluate fit between predicted and observed LPI values in one sample when equations reflected regression weights derived in another.

Findings of the Regression and Residual Analyses: Total Sample

Table 7 shows intercorrelations and descriptive statistics for designated study variables in the combined sample ($N = 738$): TOEIC scores (LC, RC, Total) in columns (a) through (c) and columns (d) through (i), the the Oralcy index, LCt (TOEFL LC mean), RCt (TOEFL Reading Comprehension and Vocabulary mean), SWEt (TOEFL Structure and Written Expression mean), TOTt (TOEFL Total mean), and a derived variable labeled *Write* (the average of RCt and SWEt, thought of as a general group-level measure for nonlistening aspects of proficiency), and finally LPI rating, in column (j). For nonlistening group-level measures (RCt, SWEt, and their average, Write), the zero-order coefficients (with LPI rating) were quite small and negative. Corresponding coefficients for the two listening-related group measures (the Oralcy index and LCt, the imputed TOEFL LC mean) were only slightly larger, but they were positive. Because the Oralcy index is a linearly derived function of the three section means, it was not included in further analyses of the data, which were concerned primarily with evaluating the contribution of the group analytic variables to prediction of LPI rating when included in sets with TOEIC scores.

Table 7

Intercorrelations and Descriptive Statistics for TOEIC Test Scores, TOEFL-Generated Group-Level Scores, and LPI Rating: Combined Sample ($N = 738$)

Variable	Intercorrelations									
	TOEIC score			Group analytic measure						LPI
	LC (a)	RC (b)	Total (c)	Oralcy (d)	LCt (e)	SWEt (f)	RCt (g)	TOTt (h)	Write (i)	rating (j)
LC	1.00	.73	.93	-.12	.10	.30	.26	.25	.28	.63
RC	.73	1.00	.93	-.25	.04	.39	.31	.21	.35	.51
Total	.93	.93	1.00	-.20	.08	.37	.31	.25	.34	.61
Oralcy	-.12	-.25	-.20	1.00	.67	-.22	-.26	.15	-.25	.19
LCt	.10	.04	.08	.67	1.00	.55	.54	.81	.55	.14
SWEt	.30	.39	.37	-.22	.55	1.00	.95	.84	.98	-.02
RCt	.26	.31	.31	-.26	.54	.95	1.00	.90	.99	-.04
TOTt	.25	.21	.25	.15	.81	.84	.90	1.00	.89	.08
Write	.28	.35	.34	-.25	.55	.98	.99	.89	1.00	-.03
LPI	.63	.51	.61	.19	.14	-.02	-.04	.08	-.03	1.00
Mean	328	301	628	40.7	49.4	50.6	50.6	49.5	50.6	1.89
SD	80	79	148	6.5	1.8	1.7	2.4	2.0	2.1	.66

Note: LCt, SWEt, RCt, and TOTt are mean section and total scores, respectively, for TOEFL examinees from the language groups here under consideration (circa 1990): LCt (Listening Comprehension); SWEt (Structure and Written Expression); RCt (Reading Comprehenion and Vocabulary); TOTt (TOEFL Total, arbitrarily divided by 10 for this study). For detail, see Tables 5 and 6, and related discussion, above.

Table 8 reports selected findings for the series of five sets of three regression and residual analyses (15 in all) that were conducted using data for the total sample. More specifically, Table 8 shows for each of the 15 analyses, beta weights for the variables involved, in columns (a) through (h), and the corresponding multiple correlation coefficient, in column (i). Columns (j), (k), and (l) provide F values, associated probabilities, and values of Eta^2, respectively, for the corresponding analyses of variance for differences among the nine study groups with respect to the residual variable associated with the indicated predictive composite. The findings summarized in Table 8 indicate that when the group-level measures representing the proficiency domains corresponding to the three TOEFL sections were included in regression equations with TOEFL scores (Series 2, 3, and 4, respectively), the multiple correlations were, in all instances, considerably larger than those obtained when only TOEIC scores were used for prediction (Series 1). The findings also showed

that, as indicated by results for Series 5, such a pattern did not obtain when TOTt (the mean TOEFL Total score) was included with TOEIC scores. At the same time, in analyses involving all three group analytic variables, only the beta weights for LCt (weighted positively) and SWEt (weighted negatively) were clearly statistically significant; the relatively much smaller (positive) beta weights for RCt tended not to meet a .05 probability criterion. It may also be seen in Table 8 that results involving LCt and Write (the RCt/SWEt composite) in combination with TOEIC scores were essentially no better than those obtained when only the dominantly weighted group-level nonlistening measure (SWEt) was included with LCt. Furthermore, it may also be seen that LCt and SWEt yielded results that appear to be fully comparable to those obtained when all three group variables were included (cf., Series 2, 3, and 4 in Table 8). In evaluating the beta weights, apart from questions of statistical significance, it is noteworthy that LCt (group-level listening comprehension) is positively weighted, whereas comparably high negative weighting is specified for SWEt (group-level knowledge of EFL structure) as well as for Write (composite of group-level nonlistening measures). This indicates a corresponding downward adjustment in predicted LPI rating for all members of a TOEIC sample from a language group for which, for example, the corresponding TOEFL reference group SWE mean (here SWEt) is higher than the corresponding TOEFL LC mean (here, LCt). Such a pattern is relatively more pronounced for native-Korean TOEFL examinees than for members of the other language groups represented in the study sample (e.g., see Table 6, above).

The net effect of the adjustments introduced by appropriately weighting the group-level measures can be inferred from the pattern of results for the corresponding sets of residual analyses. These results indicate a general pattern of improvement in degree of fit between means for group distributions of observed and predicted LPI rating when group-level measures were included in the corresponding regression equations (cf. results of one-way analysis of variance for Series 2, 3, and 4, with those for Series 1, in Table 8, below). This improvement in fit is highlighted in Table 9, which shows detailed findings of several one-way analyses of variance for which general results were reported in Table 8. More specifically, Table 9 shows (in the first data column) mean residuals (LPI rating minus predicted rating) for the nine TOEIC samples involved (treating separately four different subsamples of native-Japanese speakers) when only TOEIC Total score was used to generate predicted LPI rating (Series 1 in Table 8, below). In the remaining columns, for comparative purposes, corresponding results are shown for residual analyses when group-level

measures (LCt and SWEt, in this instance) were included, in turn, with TOEIC LC only, TOEIC LC and RC as separate predictors, and TOEIC Total, the simple sum of LC and RC (see Series 4 in Table 8, below). It is apparent that including the group analytic variables with a TOEIC score (or scores) resulted in a relatively marked decrease in values of F, and that this was especially true for the sample of native-Korean speakers (cf., mean residuals in Table 9 for TOEIC Total only, with those resulting from the introduction of group-level measures). Fit improved also for the relatively large Thai sample.

Table 8

Selected Results of Exploratory Regression and Residual Analysis Designed to Assess the Contribution of the Several Group-Level Measures to Prediction of LPI Rating in the Combined Sample ($N = 738$)

Analysis	Beta weight(s) for predictor(s)								One-way anal.of var.			
	TOEIC score			Group analytic variable					*R*		F Prob.	Eta^2
(Col.)	(a) LC	(b) RC	(c) Tot	(d) LCt	(e) SWEt	(f) RCt	(g) Write (e+f)	(h) TOTt	(i)	(j)	(k)	(l)
1a	.63								(.629)	19.3	<.0001	.234
1b	.55	.10						.633		22.8	<.0001	.200
1c			.61						(.612)	27.9	<.0001	.175
2a	.72			.28	-.41	.02*			.707	1.5	<.1600	.016
2b	.53	.31		.33	-.61	.13*			.733	1.1	<.3800	.012
2c			.77	.35	-.65	.16*			.729	1.2	<.3200	.013
3a	.71			.27			-.38		.703	2.5	<.0700	.027
3b	.53	.26		.32			-.44		.723	3.8	<.0003	.040
3c			.74	.33			-.47		.717	4.6	<.0001	.048
4a	.72			.28	-.39				.707	1.5	<.1529	.016
4b	.52	.30		.33	-.48				.732	1.4	<.1958	.015
4c			.77	.35	-.50				.728	1.6	<.1098	.018
5a	.65							-.09	.635	18.6	<.0001	.233
5b	.57	.11						-.09	.639	22.3	<.0001	.200
5c			.63					-.08	.617	27.8	<.0001	.175

Note: Series 1 analyses (1a – 1c) represent results when only TOEIC test scores were included. In Series 2 (2a – 2c), three group scores were included. Series 3 (3a – 3c) differs from Series 2 in that instead of treating the two nonlistening means (SWEt and RCt) separately, the average value of the two (called *Write*) was used. In Series 4, variance associated RCt (group-level reading comprehension) was eliminated—only LCt and SWEt (representing, respectively, group-level listening comprehension and knowledge of EFL structure and written expression) were retained And finally, in Series 5, the TOEFL Total mean (TOTt) was used as a group-level measure of general EFL proficiency. Column (i) shows coefficients of multiple correlation involving the independent variables for which beta weights are shown in the same row (the two parenthesized entries for analyses 1a and 1c are zero-order coefficients. The entries in columns (j), (k), and (l) are, respectively, values of F, the associated probabilities, and values of Eta squared, respectively, resulting from one-way analyses of variance of differences among the nine groups with respect to mean residuals resulting from use of analysis equations in the designated series (df 8,729 for each analysis).

* P > .050

Table 9

Mean Residuals (LPI Rating minus Predicted Rating) when Prediction is Based on TOEIC Total Only, and on Designated TOEIC Score(s) Augmented by TOEFL-Generated Measures for Group-Level Listening Comprehension (LCt) and Knowledge of Structure and Written Expression (SWEt), Respectively

		Predictor(s) in regression equation			
		TOEIC	TOEIC score(s) and group measures		
Group	*N*	Total only	LC LCt, SWEt	LC, RC, LCt, SWEt	Total, LCt, SWEt
France	56	-.11	.03	.03	.04
IIST84Jp	66	-.06	-.06	-.09	-.10
IIST86Jp	55	-.05	-.07	-.10	-.11
IIST87Jp	42	.05	.07	.02	-.00
TOE85Jp	122	.04	.05	.01	-.00
Saudi88	10	.44	-.35	-.22	-.16
Mexico88	42	.18	-.02	-.02	-.03
Thai Air	196	.29	.03	.05	.07
Korea89	149	-.42	-.03	-.00	.01
Total	738	.00	.00	.00	.00
F (df 8,729)		27.888	1.502	1.394	1.641
Prob.		<.0001	.153	.196	.110
Eta^2		.175	.016	.015	.018

Note: LCt is the TOEFL Listening Comprehension mean; SWEt is the mean for TOEFL Structure and Written Expression. See data for Series 1 versus Series 4 in Table 8, above, for other findings from the analyses involving TOEIC Total only versus the TOEIC scores and group-level measures designated in this table.

Cross-Validation Findings

In evaluating the total-sample results just reviewed, it should be kept in mind that the degree of fit obtained between observed and predicted LPI distributions for the respective language groups reflects best least squares fit applications of the various equations involved for same sample data. Table 10 shows illustrative results of the simulated double cross-validation model described earlier. For each of the half samples defined earlier, mean residuals by language group are shown for equations involving TOEIC scores with the LCt and SWEt, the two group analytic measures contributing most to prediction in the total sample, and illustratively for equations involving TOEIC Total. Also shown are the corresponding one-way analysis of variance results for differences among

subgroup means with respect to the corresponding residuals (values of F and the corresponding probabilities). Intercorrelations and descriptive statistics for the variables involved, analogous to those reported above (in Table 7) for the total sample, are provided in Appendix B for the half-samples, A and B, respectively. The pattern of outcomes for analyses involving other TOEIC score predictors and these group-level proficiency scores was similar to that shown here for TOEIC Total.

Comparison of the cross-validation results in Table 10 with those shown for the total sample in Table 9, above, indicates that for equations involving the group analytic measures, as expected under cross- validation conditions, overall fit between distributions of observed and estimated LPI values was not quite as good as that observed under best least squares conditions in the total-sample analysis. However, the sustained superiority of equations that reflected effects of group-level measures over those without such measures is pointed up clearly by comparing the cross-validated results in Table 10 for predictions involving TOEIC Total only (used illustratively) with the corresponding results for predictions based on TOEIC scores and the designated group-level measures. Moreover, it may be seen that use of TOEIC Total only for prediction resulted in a consistent pattern of mean residuals in both B→A and A→B applications (that is, mean residuals in both cross-validated applications) tended to be either positive or negative, indicative of a consistent predictive bias. For example, means (B→A vs. A→B) when TOEIC Total was the sole predictor were -.31 and -.53 for the Korean sample, and .38 and .21 for the Thai sample. On the other hand, the corresponding half-sample values when LCt and SWEt were included with TOEIC Total were 06. and -.06 for the Korean sample, and .13 and -.12 for the Thai sample, suggesting results more consistent with sampling variation.

Table 10

Mean LPI Residuals by Language Group for Designated Half-Samples A and B, Respectively, when Regression Equations Developed in A Were Used to Predict LPI Values in B (A→B), and vice versa (B→A), with Associated ANOVA Results

Sample	Half A *n*	Half B *n*	Mean residual for equation by order of "development→application" half-sample: Total, LCt,SWEt B→A	Total, LCt,SWEt A→B	LC,RC, LCt,SWEt B→A	LC,RC, LCt,SWEt A→B	LC, LCt,SWEt B→A	LC, LCt,SWEt A→B	Total only B→A	Total only A→B
France	29	27	- 11	.21	-.11	.17	-.07	.13	-.11	-.09
Japan	141	144	.02	-.12	.04	-.10	.08	-.08	.04	-.04
Saudi	5	5	.06	-.37	- 01	-.41	-.17	-.52	.69	.20
Mexico	20	22	.27	.18	-. 4	.19	-.23	.17	.04	.29
Thai	98	98	.17	-.03	.16	-.04	.15	-.08	.38	.21
Korea	75	74	.13	-.12	.13	-.14	.13	-.18	-.31	-.53
Total	368	370	.06	-.06	.06	-.06	.08	-.08	.06	-.06
f			4.53	4.63	3.95	4.42	3.11	4.24	27.71	19.88
p			.0005	.0004	.0017	.0006	.0092	.0009	<.0001	<.0001
df, all B→A		(5,362)								
all A→B		(5,364)								
Eta^2			.0597	.0607	.0531	.0579	.0420	.0549	.2776	.2156

Note: Data for the four different samples of native-Japanese speakers, treated separately in the total-sample analyses, were combined for the cross-validation analysis. B→A indicates that the designated predictive equation was developed in sample B and used to generate predicted LPI rating in A, and A→B indicates the opposite "development→application" pattern. LC, RC, and Total, respectively, designate the corresponding TOEIC Test scores; LCt and SWEt designate the corresponding group analytic measures involved.

DISCUSSION

The findings that have been reviewed attest implicitly to the importance of continued empirical assessment of TOEIC/LPI relationships in linguistically diverse subpopulations. Results for the Korean sample call attention to the possibility that for certain national/linguistic subpopulations, general interpretive guidelines may result in systematic overestimation or underestimation of LPI-assessed speaking proficiency. It is clearly important to examine fit between observed LPI ratings and those predicted from TOEIC scores in other samples of native speakers of Korean.

Indirect Support for the Developmental Lag Hypothesis

Apart from these considerations, the findings reviewed above tend to provide indirect support for the differential developmental lag working hypothesis that guided the analyses undertaken in Phase II of this study. It was hypothesized for working purposes that the observed overestimation of LPI performance in the Korean sample when a general sample equation involving TOEIC scores is employed would be expected if acquisition of EFL speaking proficiency (as assessed by the LPI procedure) lagged relatively more behind acquisition of the EFL skills and knowledge being tapped by the TOEIC LC and RC sections for typical native-Korean-speaking EFL learners than for demographically comparable EFL learners in the other national/linguistic settings represented in the present study sample. In an indirect assessment of this hypothesis, TOEFL section means (for LC, SWE, and RCV, respectively) for a representative array of native-language reference groups (N =110) were introduced as the primary units of analysis. The TOEFL LC mean was regressed on the two TOEFL nonlistening means (SWE and RCV), and the corresponding residual (LC mean minus predicted LC mean) was computed. The native-Korean-speaking contingent was found to have a lower LC mean than that of other native-language contingents with comparable means on the nonlistening sections. Based on assumptions and empirical evidence introduced at the outset, it was reasoned that the relative (latent) level of EFL speaking proficiency in a TOEFL reference group will not tend to exceed that group's level of EFL listening comprehension. Thus, the overestimation of TOEFL LC mean for native-Korean speakers in the TOEFL testing context was deemed to parallel conceptually the Phase I finding that the use of previously developed regression-based guidelines for inferring level of LPI-assessed EFL speaking proficiency from TOEIC scores resulted in systematic overestimation of LPI rating for native-Korean speakers in the TOEIC testing context.

The inference of conceptual parallelism appears to be reinforced by findings of exploratory multilevel analyses in which LPI rating was regressed on a set of independent variables that included TOEFL section means as scores on native-language-group analytic variables, along with individuals' TOEIC scores. In these analyses, each individual had two sets of scores, namely, his or her scores on the TOEIC Test and the set of TOEFL section means for his or her particular language group. Significant language-group-related effects were found to be associated with the group analytic variables (TOEFL section means). For example, beta weights were statistically significant for two of the three group-level scores, and the patterning of the weights was judged

(retrospectively) to be theoretically consistent with the differential developmental lag hypothesis. Thus, in total-sample regression analyses, the TOEFL mean for Listening Comprehension (here labeled LCt) was positively weighted, while the TOEFL mean for Structure and Written Expression (SWEt) was negatively weighted, after controlling for performance of individuals on the TOEIC Test. Hence, the net contribution of the group-level measures to prediction of LPI rating was a function of the direction and extent of the difference between the two TOEFL-generated group-level measures involved. An appropriate net negative adjustment was introduced in LPIs estimated for members of TOEIC samples from language groups characterized by higher SWEt than LCt. A corresponding net increment was similarly introduced for members of TOEIC samples representing language groups characterized by lower SWEt than LCt. The overall net effect of these adjustments was improvement in fit between sample distributions of observed and estimated LPI rating, especially so for the sample of native speakers of Korean, for which the group-level LCt score was relatively lower than the corresponding group-level SWEt score (cf. scores in Table 6). Results of a simulated double cross-validation of these findings, based on analyses in two halves of the current sample, were generally consistent with those in the total sample. Of course, given the apparently novel treatment of TOEFL means in the exploratory analyses involved in this study, it is important to extend such analyses to other similarly selected samples from the language groups represented in this study, to samples of EFL learners from a representative array of national/linguistic populations, and, ideally, to such samples in which concurrently observed scores on measures of speaking, listening, reading, and writing components of EFL proficiency are available for all individuals.

Suggestions for Future Research

The findings of this study lend substantial support to the working assumption that distributions of TOEFL means for native-language reference groups tend to reflect, among other things, differences in typical EFL learning rates for native speakers of the corresponding languages. However, from time to time throughout this report, a caveat about inability to control for EFL learning conditions has been attached to evaluative references to differences in average performance for TOEFL native-language reference groups. Research designed to provide information about EFL learning-context-embedded differences in learning conditions could provide information needed to permit better-informed interpretive inferences from observed reference group

means. As they pertain to differences in EFL learning rates, however, indirect inferences drawn from TOEFL means need to be buttressed by empirical studies concerned more directly with assessing differences in the relative learning difficulty of English as a foreign language for speakers of other languages.

Research on EFL Learning Conditions

Information about differences in English-language learning conditions across the diverse national/linguistic contexts is needed. It would be useful, for example, to conduct a systematic survey of EFL learning conditions in all countries that consistently provide a significant number of candidates for tests such as the TOEFL Test and the TOEIC Test. Such a survey could be designed to document the principal patterns of English-language acquisition and use within each country, national policies affecting study of English as a foreign language in the general school population, and so on.

One type of information that would be useful is suggested by the following summary description for the Japanese EFL learning context (Ito, personal communication, cited in Wilson, 1989: Endnote 16, p. 69):

> The typical Japanese university graduate has had approximately 1,000 formal classroom hours of EFL instruction . . . spread over a span of some 8 years, beginning with middle school, distributed as follows:
> Middle school: 3 hrs/week, by 35 weeks, by 3 years = 315 hrs,
> High school: 3 hrs/week, by 35 weeks, by 3 years = 525 hrs,
> University: 3 hrs/week, by 30 weeks, by 2 years = 180 hrs.

Thus, in evaluating the TOEFL means of successive contingents of, for example, graduate-school-oriented Japanese examinees, it is possible to take into account the fact that the observed means tend to reflect average outcomes of a common core of formal EFL instruction spread over eight years, plus formal and informal learning from all other sources (e.g., language schools; English-language reading, listening, speaking, and writing), plus effects associated with selection into the TOEFL-taking population. According to one recent source (California Foreign Language Curriculum Framework and Criteria Committee [CFLCFCC], 1999: p. 4), "[I]n most European countries and Japan and Korea, five to seven years generally are allocated to the study of another language." It is plausible that English is one of the foreign languages involved not only in Japan but also in the other countries to which the committee report alluded. Given comprehensive and

periodically updated information similar to that shown above, for each national EFL learning context, it would be possible to evaluate more adequately observed differences in patterns of proficiency attainment across native-language (and native country) reference groups. Such information would clearly permit better informed (and potentially more valid) interpretive inferences from operationally collected test data for samples of test takers from diverse national/linguistic populations of EFL learners. Such information plausibly could be supplied by EFL professionals representing EFL proficiency testing programs (such as the TOEFL Test or the TOEIC Test, for example) in countries throughout the world.

In the absence of such surveys, or ideally in conjunction with them, it would be useful to collect routinely information from EFL proficiency-test candidates on personal and other background variables—information that might usefully supplement the information collected about the respective national EFL/ESL learning contexts. For example, in an investigation concerned with identifying possibly useful items of background information to be supplied by TOEFL candidates, Feldmesser (1982) concluded as follows:

> From the sifting through of some two dozen variables . . . four . . . emerged as being the most promising: total number of years of formal study of English; reading of English-language newspapers, books and magazines; number of years of father's education; and number of years of mother's education. Each of them was independently related to TOEFL scores. (p. 28)[10]

Research on Comparative EFL Learning Difficulty for Native Speakers of Other Languages

That there are basic differences in the relative learning difficulty of other languages when studied as foreign languages by native-English speakers appears to be accepted as axiomatic by those concerned with foreign language teaching/learning/assessment in the United States.[11] Moreover, such differences are clearly documented in an important, empirically validated fourfold classification of some 40 other languages according to relative learning difficulty for native-English speakers. This classification was developed by the Defense Language Institute Foreign Language Center (DLIFLC), which conducts intensive foreign language programs for members of the U.S. armed services (see Lett & O'Mara, 1990).

Table 11 (adapted from Lett & O'Mara, 1990) shows languages in each of four classifications according to relative learning difficulty observed for native-English speakers. The categories are labeled, respectively, *least difficult*, *less difficult*, *more difficult,* and *most difficult*. These classifications reflect, among other things, differences in observed patterns of attainment of

Table 11

Classification of 40 Languages According to Empirically Determined Learning Difficulty for Linguistically Mature Native-English Speakers (U.S. Military Personnel): From Lett and 0'Mara (1990, p. 224)

Relative learning difficulty for linguistically mature native English-speakers

Difficulty category				
I	II	III		IV
Least difficult	Less difficult	More difficult		Most difficult
Afrikaans	German	Afghan/Dari	Hungarian	Arabic
Danish	Hindi	Albanian	Lao	Chinese
Dutch	Indonesian	Amharic	Nepalese	Japanese
French	Malay	Basque	Persian	Korean
Haitian	Romanian	Bengali	Polish	
Italian	Urdu	Bulgarian	Pushtu	
Norwegian		Burmese	Russian	
Portuguese		Cambodian	Serbo-Croatian	
Spanish		Czech	Tagalog	
Swahili		Finnish	Thai	
Swedish		Greek	Turkish	
		Hebrew	Vietnamese	

Note: The military personnel involved are judged to be demographically comparable to students enrolled in U.S. community colleges. This classification is based on observed differences in the proportion of students attaining graduation proficiency standards in programs of intensive training (6 hours per day, over 25 to 47 weeks of study). See Table 12 for a summary of illustrative differences in patterns of attainment, by language-difficulty category and scores on a foreign-language aptitude test used by the DLIFLC. For essentially the same fourfold classification of target languages, Hadley (1993) reports sharp differences in expected levels of target-language speaking proficiency (as elicited and rated according to the LPI procedure) for civilian personnel in training with the Foreign Service Institute of the U.S. Department of State.

graduation proficiency standards by students of languages in the respective difficulty categories. Characteristic differences in outcomes by difficulty category are summarized in Table 12 (adapted from Lett & O'Mara, 1990). For present purposes it is sufficient to note that the validity of the classification of other languages according to relative learning difficulty for native-English speakers appears to be strongly supported by the outcome data summarized in Table 12.[12] No such empirical classification involving English and a comparable array of other languages as target languages for foreign language learners in a non-English-dominant foreign-language learning context could be unearthed in periodic informal reviews of potentially relevant literature, combined with informal queries addressed to colleagues in other countries and to selected embassies in the United States.

Table 12

Expectancy of Meeting Graduation Standard in Designated Categories of Foreign Languages, by Designated Score Levels on the Defense Language Aptitude Battery (DLAB): Adapted from Lett and O'Mara (1990, pp. 234-237)

Relative learning difficulty of language	Percent meeting graduation standard, by score level on the DLAB			
	DLAB < 90		DLAB ge > 114	
	N	Percent	*N*	Percent
Category				
I (Least difficult)	124	44.0	80	> 63.8
II (Less difficult	46	< 28.3	73	> 70.0
III (More difficult)	none	(18.3) #	307	49.2
IV (Most difficult0	none	(6.7) ##	241	29.5

Note: Aptitude (DLAB) requirements increase with the relative learning difficulty of the language under consideration. See Table 11 for enumeration of languages included in the respective categories. Detail regarding the DLAB is not essential for present purposes (see Lett & O'Mara, 1990).

An aptitude test score of 95 or higher is required for study of a Category III-difficulty language. Accordingly, there are no students in the DLAB < 90 category. This percentage reflects the performance of students (*N*=180) scoring at least 95 on the test used.

An aptitude test score of 100 or higher is required for study of a Category IV-difficulty language. Accordingly, there are no students in the DLAB < 90 category. This percentage reflects the performance of students (*N*=119) scoring at least 100 on the test used.

Suggested Lines of Inquiry

The classification of languages in Table 11 may constitute a useful point of departure in designing needed studies concerned with documenting the relative learning difficulty of English as a foreign language for speakers of other languages. More specifically, there is reason to hypothesize that the empirically validated DLIFLC classification of languages according to relative learning difficulty for native-English speakers may tend also to constitute a valid classification of the other languages involved according to relative EFL-learning difficulty for corresponding native speakers. For example, as can be discerned in Table 13, languages classified as *least difficult* for native-English speakers in the DLIFLC setting include those for which the corresponding TOEFL native-language contingents have tended to exhibit relatively high average TOEFL scores, whereas the

Table 13

TOEFL Total Means for 1977, 1982, 1990, and 1996, respectively, for Native Language Reference Groups Selected to Represent Groups of Languages Identified by DLIFLC as Being *Most Difficult* and *Easiest*, respectively, for Native-English Speaking Language Students to Learn (see Table 11, above)

Language	Relative learning difficulty for native-English speakers #	Reported TOEFL Total mean for corresponding language contingent			
		Year of testing			
		1977	1982	1990	1996
Arabic	Most difficult	477	463	480	507
Chinese	Most difficult	506	503	509	532
Japanese#	Most difficult	483	487	485	499
Korean	Most difficult	496	504	505	518
Danish	Easiest	585	584	593	606
Dutch	Easiest	588	584	598	609
Icelandic	Easiest	581	568	563	590
Swedish	Easiest	592	582	591	591

Note: TOEFL Total means for these and other time periods are shown in Appendix A, Table A-1, for some 68 native-language reference groups.

A complete enumeration of the languages involved is provided in Table 11, above.

opposite tends to be true for the languages classified as *most difficult* for native-English speakers. More generally, results of unpublished analyses (Wilson, 1995) indicate that the underlying pattern of covariation suggested by the selected findings shown in Table 13 tends to be consistent when TOEFL means for contingents of native speakers of all of the languages represented in the respective DLIFLC categories are similarly considered (see also Appendix A, Table A-1, herein).

Moreover, it would appear that the DLIFLC classification of languages based solely on their empirically determined relative learning difficulty for native-English speakers results incidentally in a grouping of languages roughly according to degree of mutual linguistic similarity/distance. It seems clear, for example, that there is greater linguistic similarity between English and the languages classified by DLIFLC as *easiest* for native-English speakers to learn, than between English and the languages classified by DLIFLC as *most difficult* for native-English speakers. And it would appear that there is substantial linguistic distance between English and the

several *most difficult* languages for native-English speakers. Thus, by logical extension, the latter languages might be expected to be correspondingly relatively difficult for native speakers of any language that is relatively closely related linguistically to English. Finally, it would seem plausible to hypothesize that to the extent that they tend to share underlying linguistic similarities, languages within the respective DLIFLC classifications may tend to lend themselves to mutual acquisition through cross-training (see, for example, Diplomatic Language Services, 1995). It would be useful to conduct direct empirical assessments of the relative learning difficulty of English and other languages, as foreign languages, for native speakers of other languages, along lines represented by the DLIFLC model, in several national/linguistic settings selected purposefully to represent languages from the four DLIFLC categories. Of course, it may not be feasible to conduct such comprehensive studies. However, it would appear to be feasible to conduct more limited research designed to document levels of EFL proficiency attainment in comparably selected, representative samples of EFL learners in similarly selected national/linguistic settings. This might be accomplished by administering well-standardized tests such as the TOEFL or the TOEIC to samples of EFL students near the end of their secondary schooling, for example, in national/ linguistic settings that provide comparable periods of formal instruction in English as a foreign language as a universal academic requirement. If other languages are similarly required, assuming the availability of comparably scaled tests of foreign language proficiency for the languages involved, the scope of inquiry could be extended to include those languages as well. Information on typical modes of instruction, curricular emphases, and so on, could also be collected.

Models for National Surveys

One potentially useful model for national assessments designed to document attainments of EFL and/or other foreign language students is represented by Carroll's (1967) study of the attainment of U.S. college majors in French, German, Spanish, and Russian, respectively, toward the end of the senior year. Using a simple equating model with data for relatively small calibration samples from the target populations, Carroll linked distributions of scores on norm-referenced tests of basic macroskills to distributions of ratings of functional levels of proficiency on conceptually comparable, quasi-absolute scales developed by the Foreign Service Institute. (See ETS, 1982a, and Lowe, 1987, for historical perspective regarding development of conceptually comparable procedures, including quasi-absolute rating schedules, for the direct assessment of speaking,

listening, reading, and writing skills in a target language.) Given that linkage, after administering the norm-referenced measure to general national samples, Carroll was able to translate the distributions of observed scores into distributions of scores on scales reflecting attainment by language and macroskill area according to conceptually comparable, behaviorally defined scale levels. Moreover, using the responses of both teachers and students to comprehensive background questionnaires, Carroll and his associates were able to control for differences in patterns of foreign language study (*regular* vs. *irregular*, for example) and extra-curricular variables that might affect language learning (e.g., extent and nature of travel/study abroad, frequency of use of the target language, and so on). Wilson noted that in this landmark study Carroll and his collaborators

> . . . demonstrated that the interpretive power inherent in behaviorally scaled direct assessments could be harnessed—by empirical linkage rules established in samples from defined populations—to psychometrically more efficient norm-referenced measures of language macroskills, and thus be extended to the populations involved. (Wilson, 1989, p. 66)

The Carroll (1967) linkage model was adapted subsequently in a study (Hilton, Grandy, Kline, Liskin-Gasparro, with Stupak and Woodford, 1985) in which distributions of self-assessments of speaking proficiency in French and Spanish by teachers of those languages in the United States were linked to distributions of ratings of performance in LPIs. Hilton et al. used a self-assessment schedule with proficiency-level descriptions designed to parallel those developed for use by professional interviewers/raters in rating LPIs. The self-assessments were then calibrated (by the method of equal standard deviations) to the distributions of LPI ratings by professional interviewers/raters in scattered, small calibration samples of teachers of the respective languages. The easy-to-administer self-assessment schedule was then used to collect information from corresponding national samples.

As part of a national survey (Barrows, 1981) concerned with assessing the range of awareness of other cultures among U.S. college students, Clark (1981) designed and reported the results of aspects of the survey concerned with students' foreign language attainments. More specifically, Clark was concerned with identifying the foreign languages in which students deemed themselves to be "most proficient" and with obtaining respondents' self-assessments of listening, reading, and speaking proficiency in the respective languages so identified. Clark used self-assessment inventories composed of so-called "can-do" items, which require students to indicate the relative ease/difficulty with which they can perform (i.e., can do) specified linguistic tasks using

a target language. The range of difficulty of individual tasks (can-do items) was judged to represent a range of difficulty roughly paralleling that of quasi-absolute native-speaker-referenced scales developed by the Foreign Service Institute for rating proficiency in basic target-language skills. A national survey (report in progress) focusing on the attainments of foreign language majors in the United States has recently been conducted under auspices of the DLIFLC (Lett, personal communication, January 2001). This survey involved the use of can-do self-assessment questionnaires composed of items that were graded (judgmentally by experts) according to the estimated likelihood that the linguistic activity specified can be performed easily by a specified majority of target language users/ learners at defined levels of proficiency on behaviorally defined rating scales.

Generally speaking, there is reason to believe that surveys of proficiency in foreign languages involving the use of appropriately designed self-assessment inventories are likely to yield findings that will permit useful, pragmatically valid inferences about comparative levels of development of basic aspects of proficiency in foreign languages in, among other things, samples of students nearing the end of their secondary school programs (e.g., Hilton et al., 1985; Ingram, 1985; LeBlanc & Painchaud, 1985; Oscarson, [in press]; Wilson & Lindsey, 1999).

REFERENCES

Angoff, W. H., & Sharon, A.T. (1971). Comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges. *TESOL Quarterly*, *5*, 129-136.

Barrows, T. S. (Ed.). (1981). *College students' knowledge and beliefs: A survey of global understanding—The final report of the Global Understanding Project*. New Rochelle, NY: Change Magazine Press.

California Foreign Language Curriculum Framework and Criteria Committee. (1999). *Draft foreign language curriculum framework K-12*. Sacramento, CA: Curriculum Frameworks and Instructional Resources Office.

Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college, *Foreign Language Annals*, *1*, 131-151.

The Chauncey Group International. (1996). *TOEIC: Report on test-takers worldwide, 1996*. Princeton, NJ: Author.

The Chauncey Group International. (1999). *TOEIC user guide*. Princeton, NJ: Author.

Clark, J. L. D. (1976). *The performance of native speakers of English on the Test of English as a Foreign Language* (TOEFL Research Report No. 1). Princeton, NJ: Educational Testing Service.

Clark, J. L. D. (1981*).* Survey measures: Language. In T. S. Barrows, *College students' knowledge and beliefs: A survey of global understanding—The final report of the Global Understanding Project* (pp. 25-35). New Rochelle, NY: Change Magazine Press.

Diplomatic Language Services, Inc. (1995, Fall). Language classification and cross training *The DLS Diplomatic Courier*. Retrieved January 15, 2001 from the World Wide Web: http://www.dls-inc.com

Educational Testing Service. (1973). *Manual for TOEFL score recipients*, *1973 edition*. Princeton, NJ: Author.

Educational Testing Service. (1978). *TOEFL test and score manual*. Princeton, NJ: Author.

Educational Testing Service. (1981). *TOEFL test and score manual*. Princeton, NJ: Author.

Educational Testing Service. (1982a). *ETS oral proficiency testing manual*. Princeton, NJ: Author.

Educational Testing Service. (1982b). *TOEFL test and score manual*. Princeton, NJ: Author.

Educational Testing Service. (1985a). *Test of English for International Communication: Bulletin of information*. Princeton, NJ: Author.

Educational Testing Service. (1985b). *TOEFL test and score manual*. Princeton, NJ: Author.

Educational Testing Service. (1986). *Guide for TOEIC users*. Princeton, NJ: Author.

Educational Testing Service. (1990). *TOEFL test and score manual*. Princeton, NJ: Author.

Educational Testing Service. (1991). *Test of Spoken English*. Princeton, NJ: Author.

Educational Testing Service. (1996). *TOEFL test and score data summary*. Princeton, NJ: Author.

Educational Testing Service. (1997). *TOEFL test and score manual*. Princeton, NJ: Author.

Feldmesser, R. A. (1982). *An inquiry into possible new items of background information about TOEFL candidates*. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Hadley, A. O. (1993). *Teaching language in context*. Boston, MA: Heinle and Heinle Publishers.

Hemingway, M. (1999). *English proficiency tests: A comparative study*. Princeton, NJ: The Chauncey Group International.

Hilton, T. L., Grandy, J., Kline, R. G., & Liskin-Gasparro, J. E., in collaboration with Stupak, S. and Woodford, P.E. (1985). *The oral language proficiency of teachers in the United States in the 1980's—An empirical study*. Princeton, NJ: Educational Testing Service.

Ingram, D. E. (1985). Assessing proficiency: An overview of some aspects of testing. In K. Hyltenstam & M. Pienemann (Eds.), *Modeling and assessing second language acquisition* (pp. 215–276). San Diego, CA: College-Hill Press.

Johnson, D. C. (1977). The TOEFL and domestic students: Conclusively inappropriate, *TESOL Quarterly*, *11*, 79–86.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument, *TESOL Quarterly*, *19*, 673–687.

Lett, J. A., Jr., & O'Mara, F. E. (1990). Predictors of success in an intensive foreign language learning context: Correlates of language learning at the Defense Language Institute Foreign Language Center. In T.S. Parry & C.W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 222–260). Englewood Cliffs, NJ: Prentice Hall Regents.

Lowe, P. L., Jr. (1987). *Interagency language roundtable oral proficiency interview*. In C.J. Alderson, K.J. Khrahnke, & C.W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 81–83). Washington, DC: Teachers of English to Speakers of Other Languages.

Lowe, P. L., Jr., & Stansfield, C. W. (1988). Introduction. In P.L. Lowe, Jr., & C.W. Stansfield (Eds.), *Second language proficiency assessment: Current issues* (pp. 1–11). Englewood Cliffs, NJ: Prentice Hall Regents.

Oscarson, M. (in press). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Language testing and assessment*: *Vol. 7. Encyclopedia of language and education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pedzahur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Orlando, FL: Harcourt Brace College Publishers.

Perkins, K. (1987). Test of English for International Communication. In J.C. Alderson, K.J. Krahnke, & C.W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 81–83). Washington, DC: Teachers of English to Speakers of Other Languages.

Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature*, *18*, 65–85.

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

Wilson, K. M. (1984). *The relationship of GRE General Test item-type part scores to undergraduate grades* (GRE Board Professional Report No. 81-22P, and ETS Research Report 84-38). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1986). *The relationship of scores based on GRE General Test item types to undergraduate grades: An exploratory study for selected subgroups* (GRE Board Professional Report No. 83-19P, and ETS Research Report 86-37). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1987). *Patterns of test-taking and score change for examinees who repeat the Test of English as a Foreign Language* (TOEFL Research Report No. 22, and ETS Research Report 87-3). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1989). *Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context* (TOEIC Research Report No. 1, and ETS Research Report 89-39). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1995). *Variation in EFL acquisition rate for native-speakers of Korean and other languages*. Unpublished report. Princeton, NJ: Educational Testing Service.

Wilson, K. M., Berquist, A., & Bell, I. (1998). *Guidelines for comparing performance on two tests of ESL proficiency: The TOEIC test and the TOEFL*. Unpublished report. Princeton, NJ: Educational Testing Service.

Wilson, K. M., & Chavanich, K. (1989). *Further evidence of TOEIC/LPI stability across diverse samples*. Unpublished report. Princeton, NJ: Educational Testing Service.

Wilson, K. M., and Lindsey, R. (1999). *Validity of global self-ratings of ESL speaking proficiency based on an FSI/ILR-referenced scale* (ETS Research Report 99-13). Princeton, NJ: Educational Testing Service.

Woodford, P. E. (1982). The test of English for International Communication (TOEIC). In C. Brumfit (Ed.), *English for international communication* (pp. 61–72). New York: Pergamon Press.

APPENDIX A

Stability of TOEFL Means and the Oralcy Index Derived from Corresponding Section Means for TOEFL Native-Language Reference Groups

The ad hoc analyses reported in this appendix were prompted by findings, reported in the text, suggesting that observed differences in levels and patterns of performance for native-language reference groups on the TOEFLTest tend to convey information that may permit valid inferences about corresponding differences in EFL learning rates (in the proficiency domains tapped by the TOEFL Test) for native speakers of the respective languages who are likely to take a test such as the TOEFL Test. The TOEFL Test is widely used to assess English proficiency in samples of international students planning to study in the U.S. or other English-speaking educational contexts (see, for example, ETS, 1973, 1978, 1982b, 1990, 1997).

The traditional paper-and-pencil version of the TOEFL Test here under consideration (see, for example, ETS, 1997, for information regarding the revised, computerized version of the TOEFL Test now being used in certain regions) has three separately timed sections labeled, respectively, Listening Comprehension (LC), Structure and Written Expression (SWE), and Reading Comprehension and Vocabulary (RCV). Scaled scores on the respective sections range between, roughly, 20 and 67; the Total score is a weighted combination of the three scaled section scores, and the corresponding score range is between 200 and 650. TOEFL total and section means for examinees classified by native language and native country, respectively, have been reported periodically by the TOEFL program for two decades or more (e.g., ETS, 1973, 1982b, 1990, 1996).

There are marked differences by native language, as well as by native country, in average performance on the TOEFL Test. For present purposes, attention is focused on the properties of the means for examinees classified solely by reported native language. In presenting average performance data for such reference groups, the TOEFL program consistently has emphasized the types of inferences that should not be drawn from the observed differences among reference groups defined by native language or country of origin, along lines suggested by the following excerpt from the 1973 *Manual for TOEFL Score Recipients* (ETS, 1973):

> It is important to point out that the data do not permit the generalization that there are fundamental differences in the *ability* of various national and language groups to learn English or in the level of English proficiency generally attained by them. . . . Because different selective factors operate in different parts of the world to determine who takes TOEFL, the samples on which the tables are based are not at

> all representative of the student populations from which they came . . . (emphasis added). (p. 15)

Such a disclaimer emphasizes the point that the relative standing of examinee-contingents by native language and/or country with respect to average TOEFL performance should not be assumed to reflect differences in ability, per se, of the corresponding national and language groups to learn English. The disclaimer also emphasizes that in evaluating differences in means for various native-language reference groups, one should not assume comparability in conditions of English-language acquisition/learning and use across the diverse national/linguistic contexts involved—conditions such as quality, intensity, duration, and age of initiation of EFL instruction; characteristic degree of exposure to native-English speakers and/or use of English as an official language, for example.

Moreover, factors influencing selection into the respective TOEFL-taking (reference group) samples may differ (varying mixes of individual initiative and institutional influence, both academic and governmental, may be involved, for example). While recognizing these limitations, it seems important, at the same time, not to overlook the plausible hypothesis that EFL-learning difficulty may tend to vary significantly across language groups. The difficulty of the task of learning English as a foreign language, per se, might vary as a function of, among other things, linguistic distance between English and the native language of prospective learners in typical EFL learning contexts.

In any event, research is needed to document and assess the interpretive implications of differences by native language and/or native country (within which native language tends to be nested) with respect to potential explanatory variables. These include, for example, proportion of the general school population participating in formal EFL/ESL study, characteristic patterns of EFL/ESL acquisition and use, factors governing selection of nationals into the TOEFL-taking population, and so on. And, of course, it would be useful to conduct studies designed to assess empirically, under controlled conditions, rates of EFL/ESL learning for speakers of a representative array of other languages. However, in the absence of such studies, it appears to be useful to conduct analyses designed to shed light on the extent to which observed differences in TOEFL means for native-language contingents tend to exhibit properties that necessarily must obtain if the periodically published means by native language are to be thought of as having more than temporally constrained descriptive value.

One necessary condition for the latter, of course, is that the observed differences must be stable over time. Two aspects of stability are of interest. In the first place, we are interested in the extent of stability in rank order of the respective native-language contingents over time, as indexed by trends in level of correlation between means of contingents across successive time periods. In the second place, it is of interest to evaluate degree of stability or change in the level of proficiency typically attained by native-language contingents, as indexed by net change over time in the means of distributions of means by native language or country exhibited by successive contingents of TOEFL-takers over time.

The only formal assessment of the correlation stability of TOEFL reference-group means that the writer could locate in the course of this investigation involved means by native country for some 100 country-contingents in the late 1970s (Wilson, 1982, p. 64). A correlation of .94 was observed between TOEFL Total means (for TOEFL-takers reporting "degree plans" only) by native country (within which native language tends to be nested) for two successive (nonoverlapping but temporally proximate) testing periods. No comparative evaluation of distributions of language-group means by time period could be located. However, given increasing emphasis worldwide on English-language acquisition and use, it is plausible that there may have been some increase over the past 25 or 30 years, say, in the amount and quality of EFL instructional resources available to prospective TOEFL-takers in many national settings.

Hence, it is plausible that average TOEFL scores for today's national/linguistic contingents might tend to be higher than those of corresponding contingents tested 30 or more years ago, say, during 1965–71. No research concerned with the possibility of systematic change at the mean, upward or downward, for distributions of average TOEFL scores by either native country or native language could be located in preparing this report. Some evidence bearing on stability/change in the TOEFL scores of individuals over different time periods is available for examinees who repeated the TOEFL Test (e.g., Wilson, 1987) after intervals ranging from a month or so, up to five or more years.

Analysis of Stability of TOEFL Means by Native Language

Using published TOEFL language group means for several arbitrarily selected, nonoverlapping testing periods as units of analysis, it was possible to assess directly questions regarding the stability of the observed differences. The particular reference groups involved in the

analysis were those represented by 25 or more cases in each of five testing periods between, roughly, 1965–71 and 1996 (see ETS, 1973, 1978, 1982b, 1990, and 1996, respectively, for detail on all reference groups for which data were reported). Data for the 1965–1971 testing period were for the five-part TOEFL Test "for foreign students seeking admission . . . who took the TOEFL from October 1966 through June 1971" (ETS, 1973, p. 24). While these data were not included in all phases of the stability analyses reported below, language groups represented in testing during 1965–1971 by more the 25 examinees were used to anchor the process of identifying groups with a record of consistent participation at the arbitrarily determined minimum level of 25 cases.

A total of 68 groups was thus identified. The year designations (e.g., "1977, 1982, . . . , 1996") tend to reflect the last year in a two-year reporting period. The 68 groups involved and the corresponding reference group TOEFL Total means for the five selected testing periods between 1971 and 1996, inclusive, are shown in Table A-1. Formal TOEFL codes for these groups are also included in Table A-1 for reference purposes. Data are presented for groups listed in descending order in terms of the 1971 TOEFL Total mean.

Descriptive statistics and intercorrelations were computed not only for TOEFL Total means for reporting years 1977, 1982, 1990, and 1996, respectively, as shown in Table A-1, below, for the 68 reference groups, but also for the three corresponding TOEFL section means (Listening Comprehension or LC, Structure and Written Expression or SWE, and Reading Comprehension and Vocabulary or RCV), respectively (not shown in the table). Salient findings are reported in Table A-2. In these analyses, means were not weighted by sample size. This was done in order to focus attention squarely on degree of stability or change in group performance, per se. Thus, the intercorrelations in Table A-2 provide systematically derived perspective on rank-order stability of language-contingent means over a 20-year period. In addition, the descriptive statistics indicate trends in central tendency and variability for the corresponding distributions of means between, roughly, 1977 and 1996, inclusive.

For present purposes, it is sufficient to call attention to the fact that the generally relatively large correlation coefficients in Table A-2 exhibit a gradually declining gradient as the time between observations increases. This pattern suggests substantial stability accompanied by gradual change in relative standing for these reference groups between 1977 and 1996—true for TOEFL Total and each of the three TOEFL sections as well. For example, coefficients reflecting stability in relative standing across adjacent periods exceeded the .90 level; for nonadjacent time periods,

coefficients tended to reach or exceed the .80 level. Only one coefficient in the table did not reach the .80 level, namely, that reflecting 1977 to 1996 stability for differences in LC means (r = .788). The means of the corresponding distributions of means exhibit a rising gradient. For each TOEFL section as well as for TOEFL Total, the mean of the distribution of means in 1996 was larger than the corresponding mean in 1977, with some differences in magnitude according to TOEFL section. Considering only the three section means, for example, the largest net increase in average tested proficiency between 1977 and 1996—as reflected in the difference between the corresponding means in 1977 standard deviation units—was on Structure and Written Expression, the second largest gain (0.52 SD units) was registered on Reading Comprehension and Vocabulary, while the least gain (0.44 SD units) was for Listening Comprehension. Such a pattern of net gain plausibly could be due, at least in part, to general improvement across learning contexts in conditions affecting ESL/EFL learning/acquisition between the mid-1970s and 1996. For example, increased opportunity for and/or improvement in the quality of formal EFL study across national/linguistic settings is suggested by the fact that change in average performance was greatest on Structure and Written Expression, which taps knowledge of the type typically emphasized in formal EFL instruction (e.g., Saegusa, 1985). And the smaller net gain for Listening Comprehension, for example, tends to be consistent with the likelihood that under typical learning conditions in academic, EFL instructional settings worldwide, aural/oral aspects of proficiency typically may be given less emphasis and/or may tend to be less readily amenable to improvement under available instructional conditions, than are aspects being tapped especially by Structure and Written Expression and, to probably a lesser extent, Reading Comprehension and Vocabulary.

In any event, the overall stability of the language group means here under consideration over more than two decades appears to exceed the stability of the TOEFL Total scores of individual test takers who repeated the TOEFL Test one or more times after periods ranging from a month or two to over five years. For example, Table A-3 (from Wilson, 1987) shows trends in correlation involving TOEFL Total scores for examinees identified as having repeated the test, classified according to total number of times tested (2, 3, . . , 9). Also provided are average elapsed times between the first and last test administrations, which varied over a 15-month range: from nine months (for one-time repeaters), some 13 months for two-time repeaters, and so on, up to about 24 months for nine-time repeaters. Descriptive statistics are shown separately for each administration in each times-tested pattern. These data provide some perspective on the stability of language-group

means as indexed by trends in correlation between repeated observations over time, as compared to the stability of individual test performance, similarly indexed, without regard to language.

Stability of the Oralcy Index

The findings shown in Table A-2, below, suggest substantial stability for language-group means, per se. In the text, in analyses involving only means reported for the 1990 language groups, a residual variable labeled *Oralcy* was derived to reflect the difference between the observed LC mean and a predicted LC value, reflecting the regression of LC mean on the two nonlistening means (SWE and RCV). To the extent that the means involved reflect stable differences among the corresponding language groups, it might be expected that regression-based indices such as Oralcy would also tend to be relatively stable—that is, for example, that differences among language groups in rates of EFL aural development relative to rates of EFL reading/writing development tend to be relatively stable. To assess this assumption, Oralcy indices were computed for each of the time periods here under consideration for language groups represented in each of the four time periods, beginning with 1977. Table A-4 shows intercorrelations of the respective indices and corresponding descriptive statistics. Stability with gradual change is indicated for this index, consistent with findings for the means involved as reported below in Table A-2. Gradual narrowing of differences among groups over time is suggested by the declining gradient in the observed standard deviations of distributions of the residual index.

Appendix A

Table A-1

TOEFL Total Means by Native-Language Reference Groups, Represented by $N > 24$ in Designated Testing Periods Between 1971 and 1996, Inclusive: Groups Listed in Descending Order by 1971 Mean

Language	TOEFL code	Mean TOEFL Total score Year				
		1971	1977	1982	1990	1996
Swedish	481	580	592	582	591	591
Dutch	419	579	588	584	598	609
Iceland	447	578	581	568	563	590
Danish	416	576	585	584	593	606
German	437	569	587	575	586	593
Sinhali#	361	561	563	545	553	545
Tamil	370	561	569	569	578	586
Ewe#	119	560	543	556	551	549
Norwegian	456	549	572	559	573	590
Ibo	136	548	515	506	532	566
Shona	170	548	559	553	583	587
Tagalog	367	548	531	551	573	575
French	434	539	552	546	551	555
Finnish	428	538	568	564	584	594
Akan	104	537	569	560	557	561
Ga	125	537	551	568	569	578
Marathi	348	535	559	569	590	601
Punjabi	355	534	536	542	540	553
Sindhi	360	533	557	564	554	555
Kashmiri	338	532	561	559	582	595
Mende	162	531	510	517	537	534
Assamese	301	528	544	523	570	540
Burmese#	307	528	517	509	513	506
Kannada	335	526	545	556	577	582
Hungarian	443	526	531	547	558	567
Nepali#	351	525	530	522	535	524
Russian	467	525	533	526	540	549
Bengali#	305	524	531	507	512	516
Kikuyu	142	523	532	550	563	564
Swahili	176	522	528	539	542	547
Malayalam	346	522	511	564	579	593
Romanian	464	521	529	542	564	581
Yoruba	188	519	513	511	548	569
Armenian	401	517	500	506	530	546
Hausa	133	515	511	511	526	546
Hebrew	507	515	539	534	551	576
Efik	116	514	525	507	529	560
Hindi	323	513	546	562	587	591

(*table continues*)

Table A-1—*Continued*

Language	TOEFL code	Mean TOEFL Total score, Year				
		1971	1977	1982	1990	1996
Telugu	373	511	536	540	553	557
Turkish	484	511	502	499	516	531
Oriya	353	507	521	534	556	571
Ilocano	326	505	504	515	530	537
Polish	459	499	523	529	547	559
Portuguese	461	499	517	515	534	551
Amharic	107	498	496	486	500	520
Czech	413	494	566	552	558	570
Chinese	315	489	506	503	509	532
Greek	440	489	503	496	526	541
Ponapean	613	484	459	464	511	530
Slovak	473	481	567	525	558	565
Somali	173	480	484	477	484	507
Spanish	478	478	521	504	534	551
Serbo-Croatian	470	477	511	527	551	565
Indonesian	328	476	490	479	496	510
Korean	340	475	496	504	505	518
Japanese	331	470	483	487	485	499
Gujarati	320	463	528	533	540	546
Samoan	616	463	418	438	516	520
Kusaiean	604	459	413	427	454	508
Vietnamese	388	458	497	499	513	504
Palauan	610	458	424	456	482	507
Lao	343	453	455	456	491	492
Farsi	504	452	452	484	509	533
Pushto	357	450	507	504	504	523
Arabic	501	450	477	463	480	507
Thai	376	446	464	473	489	494
Marshalles	607	445	393	410	451	462
Trukese	625	438	413	420	443	479

Note: The TOEFL Total means are as reported by the TOEFL program for 68 native-language groups (except Italian) that were represented by 25 or more cases in each of four testing periods between 1971 and 1996, inclusive, designated here by the last year of the (typically) two-year periods for which summary statistics traditionally have been reported by the TOEFL program (see ETS, 1973, 1978, 1982b, 1990, and 1996, respectively, for complete reference group data). Data for native-Italian speakers were not included due to changes in reference-group composition across periods (internal communication from the TOEFL program, March 10, 1996). The TOEFL native-language code is included for reference purposes. Means for 1971 reflect performance on the five-part TOEFL Test. They are pertinent because the TOEFL program took steps to maintain general scaling equivalence for the TOEFL Total score in the transition from the five-part to the three-part, paper-and-pencil version of the TOEFL Test. In any event, a general upward trend in these means is evident. The trend appears to be more pronounced for some native-language contingents than for others.

Language groups for which the 1996 mean is lower than that for 1971. For all others, the opposite is true.

Appendix A
Table A-2

Descriptive Statistics and Stability Coefficients for TOEFL Means of 68 Native-Language Groups for Designated Testing Periods Between 1977 and 1996[#]

TOEFL test	TOEFL testing date 1977	1982	1990	1996	Mean of means	SD means	Net change in mean of means, 1977 to 1996 (in '77 SD units)##
Total							
1977	1.000	.945	.882	.822	519.7	44.7	
1982	.945	1.000	.940	.887	521.1	41.0	
1990	.882	.940	1.000	.942	538.0	36.5	
1996	.822	.887	.942	1.000	549.4	33.5	+0.66
LC							
1977	1.000	.936	.854	.788	52.8	4.3	
1982	.936	1.000	.908	.853	52.6	4.0	
1990	.854	.908	1.000	.943	54.1	3.6	
1996	.788	.853	.943	1.000	54.7	3.3	+0.44
SWE							
1977	1.000	.945	.878	.805	51.2	5.0	
1982	.945	1.000	.943	.872	51.8	4.6	
1990	.878	.943	1.000	.930	54.2	3.9	
1996	.805	.872	.930	1.000	55.6	3.9	+0.88
RC&V							
1977	1.000	.952	.902	.849	51.9	5.2	
1982	.952	1.000	.951	.894	52.1	4.6	
1990	.902	.951	1.000	.924	53.3	4.1	
1996	.849	.894	.924	1.000	54.6	3.6	+0.52

[#] The descriptive statistics and correlations tabled are for TOEFL means as reported by the TOEFL program for 68 native-language groups, 1977, 1982b, 1990, and 1996, respectively. The 68 groups were those represented by 25 or more cases in each of five testing periods between 1971 and 1996, inclusive, as reported by ETS (1973, 1978, 1982b, 1990, and 1996, respectively). See note to Table A-1, above, and related text for additional detail. Each designated testing year refers to the last year of a two-year period.

[##] These are means of 68 means, unweighted by sample size. Means for all examinees tend to be lower. For example, some large language-group contingents typically earn relatively low average scores; also, reference group data include multiple test scores for typically lower-scoring candidates who repeat the TOEFL Test. Some years ago, it was reported (Wilson, 1987, p. S-11) that although repeaters accounted for only 28% of all test takers, they generated more than 50% of all test records in program files for a designated time period.

Appendix A

Table A-3

Trends in Correlation Between Time-1 and Time-T Total Scores, by Number of Times Tested, for Repeaters Without Regard to Analysis Group, with Corresponding Descriptive Statistics (from Wilson, 1987, Table 12, p. 34)

Correlation between t-1 score and score for designated "last" administration (t1, tT2, t1, tT3, . . , t1, tT9)

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
Mdn *r*		.776	.732	.698	.687	.678	.632	.664	.552	

Times tested	Means for designated administrations 1st	2nd	3rd	4th	5th	6th	7th	8th	9th
(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

Times Tested	Corresponding standard deviations 1st	2nd	3rd	4th	5th	6th	7th	8th	9th
(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: *N*s for means and standard deviations are as shown for correlations. These analyses are based on a 20% sample of repeaters.

[#] Mean interval in months between designated administrations, t-1 to t-T. For example, for examinees tested only two times, the t-1 to t-T interval was 9.0 months.

Appendix A
Table A-4

Stability Coefficients and Descriptive Statistics for Oralcy Indices circa 1977, 1982b, 1990, and 1996, Respectively, for 83 TOEFL Native-Language Contingents Represented in All Four Time Periods[#]

Derived variable	Derived variable					
	Oralcy77	Oralcy82	Oralcy90	Oralcy96	Mean	SD
	r	*r*	*r*	*r*		
Oralcy77	1.00	.94	.85	.74	.00	2.89
Oralcy82	.94	1.00	.89	.79	.00	2.62
Oralcy90	.85	.89	1.00	.88	.00	2.48
Oralcy96	.74	.79	.88	1.00	.00	1.80

Note: With language-group means as units of analysis, the respective Oralcy indices are residual variables defined as "LC minus predicted LC," where the predicted value reflects the regression of mean LC on mean SWE and mean RCV in data for a given time period using data generated in operational TOEFL test administrations (see ETS, 1978, 1982b, 1990, and 1996, respectively, for reported means).

[#]Although native-Italian speakers were represented in all four time periods, corresponding data were not included in this analysis. According to information provided by the TOEFL Program (internal communication, March 10, 1996, due to a significant change in policies affecting selection of TOEFL International and Special Center Program candidates in Italy, means reported for Italian candidates subsequent to the change (e.g., for the most recent of the four periods selected for this analysis) are not comparable to those reported before the change (e.g., for the three earlier periods here under consideration).

Appendix B

Ancillary Findings of Cross-Validation Analysis

As indicated in the text, two validation samples —labeled "A" and "B"—were defined by arbitrarily treating as different samples the first n/2 and the second n/2 cases, respectively, from each of the six language groups represented in the total study sample. Where 1 represents the first n/2 cases and 2 represents the second n/2 cases in the study file for each language group, the pattern followed in forming the combined half-samples, A and B, was as indicated below:

Half-	Samples, by language												
Sample	French		Japanese		Arabic		Spanish		Thai		Korean		*N*
A =	1	+	2	+	1	+	2	+	1	+	2	=	368
B =	2	+	1	+	2	+	1	+	2	+	1	=	370

Intercorrelations and descriptive statistics for designated study variables in the two samples thus defined are provided in Table B-1.

Appendix B
Table B-1

Table B-1

Interrcorrelations and Descriptive Statistics for Two Half-Samples, A (N = 368) and B (N = 370)

Half-sample Variable	LPI Rating	Group Score LCt	Group Score SWEt	TOEIC Test Score List	TOEIC Test Score Read	TOEIC Test Score Total	Descriptive Statistics Mean	Descriptive Statistics SD
			Intercorrelations					
Half A/variable								
LPI	1.000	.062	-.082	.659	.518	.635	1.9	.7
LCt	.062	1.000	.558	.097	.055	.082	49.5	1.8
SWEt	-.082	.558	1.000	.255	.384	.342	50.6	1.7
List	.659	.097	.255	1.000	.725	.931	328.1	82.9
Read	.518	.055	.384	.725	1.000	.926	303.6	79.7
Total	.635	.082	.342	.931	.926	1.000	631.7	151.0
Half B/ variable								
LPI	1.000	.218	.042	.596	.497	.585	1.9	.6
LCt	.218	1.000	.546	.108	.029	.073	49.5	1.8
SWEt	.042	.546	1.000	.352	.388	.397	50.6	1.7
List	.596	.108	.352	1.000	.740	.930	327.7	76.1
Read	.497	.029	.388	.740	1.000	.935	297.1	79.0
Total	.585	.073	.397	.930	.935	1.000	624.8	144.6

Note: The variables are as follows:

LPI rating = Language Proficiency Interview rating
LCt = TOEFL Listening Comprehension mean for corresponding native-language group
SWEt = TOEFL Structure and Written Expression mean for corresponding native-language group
List = TOEIC Listening Comprehension score
Read = TOEIC Reading Comprehension score
Total = TOEIC Total score

NOTES

1. The sample of Arabic speakers was quite small (*N*=10), but the pattern of outcomes was generally similar to those for the other language groups represented.

2. The fact that the observed correlation coefficient for one of two relatively closely related test sections with an external criterion is equal to that for the total test score, which is the sum of scores on the two sections involved, is in and of itself noteworthy quite without regard to the magnitude of the observed difference, and becomes more so if there are repeated instances in which the same pattern emerges. For consideration of such a pattern involving relationships of part versus total scores on college- and graduate-level admission tests to grade-average criteria, see Wilson (1984, 1986). In the present context, the correlation pattern noted appears to be quite consistent with the functional linkage between development of listening and speaking abilities. This linkage tends to be reflected in LC/PLI coefficients that are higher than the RC/LPI coefficients, and in LC/LPI and Total/LPI coefficients that are comparable.

3. Based on evidence that very few individuals scoring below 500 on the TOEIC Test could be expected to meet program-specified LPI levels, only trainees with Total scores at or above this level were interviewed in the Thai Air sample. See Wilson and Chavanich (1989) for a detailed description of that sample.

4. During his tenure at ETS, the interviewer/rater periodically conducted workshops concerned in part with assuring that professionally trained interviewers/raters in scattered TOEIC assessment sites remained "calibrated" with respect to a standard defined by recorded sets of previously rated interview protocols; and was himself similarly "recalibrated" periodically. Moreover, he had generated LPI ratings for three of the four linguistic samples included in the basic TOEIC calibration study, namely, the French, Mexican, and Saudi samples. And, based on his experience in Korea, it was his judgment that the sample under consideration was generally representative of the population of educated EFL learners/users who tend to take the TOEIC Test in Korea.

5. It is assumed a priori that under usual conditions of foreign language learning, development of speaking proficiency in a target language tends to lag behind development of other aspects of proficiency, including listening comprehension. "Usual conditions of foreign language learning" is intended here to (a) connote formal exposure to curriculum-embedded, academic programs of instruction in a target language (English as a foreign language, in this instance) augmented by informal, experiential learning on an individual basis, and (b) involve target-language learners who have attained linguistic maturity in the mother tongue (that is, who are some 12-13 years of age or older when formal foreign language instruction is initiated). There is limited, albeit relatively strong, empirical evidence suggesting that this assumption tends to be tenable. Carroll (1967) assessed the attainment of U.S. college foreign-language majors nearing the end of their senior year. Using procedures too detailed for brief summarization here, Carroll (1967) generated estimates of levels of listening, speaking, reading, and writing proficiency, respectively, for U.S. college seniors majoring in French, German, Spanish, and Russian, respectively, nearing the end of their college senior year. He used ratings on conceptually comparable quasi-absolute scales—then only relatively recently developed by the Foreign Service Institute's Language School—for eliciting and rating samples of linguistic behavior in the respective macroskill domains. For each domain, the rating scales involved six basic points or levels describing performance ranging from Level 0 (*no proficiency*) through Level 5 (*proficiency comparable to that of an educated speaker*). For working purposes, assuming the functional comparability of scale levels across proficiency domains and languages, Carroll concluded that for students of each target language, average levels of speaking proficiency were lower than average levels of listening comprehension, reading ability, and writing ability. No comparable study appears to have been conducted since that time. However, studies comparing the performance of native and nonnative English-speaking students on the TOEFL Test (e.g., Angoff & Sharon, 1971; Clark, 1976; Johnson, 1977) provide indirect evidence that the pattern reported by Carroll (1967) for native English-speaking majors in several foreign languages tends also to be obtained in samples made up of international students who have studied English as a foreign language (EFL)—that is, TOEFL examinees. Angoff and Sharon (1971), for example, reported percentile distributions and summary statistics for a sample of U.S. college freshmen and TOEFL candidates, respectively, on sections of the original five-part TOEFL Test. The five-part test had sections labeled, respectively, Listening Comprehension, English Structure, Vocabulary, Reading Comprehension, and Writing Ability

(indirectly assessed). Distributions of section scores were arbitrarily scaled so that each scale would have a mean centered around 50, but it was noted (see, for example, ETS, 1973, p. 13) that this figure was " . . . selected simply for convenience in score reporting . . . [and that] *a reported part score of 50 on any form of the TOEFL has no absolute meaning with respect to level of proficiency in English*" (emphasis added). However, when TOEFL general reference group part-score means (of approximately 50 on each part) were referenced to the distributions of part scores for the U.S. college-freshman sample, the degree of overlap between native- and nonnative-speaker distributions varied markedly according to proficiency domain. When these means were z-scaled relative to parameters for the U.S. sample, results were as shown below. Means on the TOEFL sections and TOEFL Total for the TOEFL examinee population are expressed as deviations from the U.S. mean in U.S. standard deviation (SD) units (adapted from data in Angoff & Sharon, 1971, Table 2, p. 131):

	TOEFL examinee mean	
Five-part TOEFL	Scaled Score	Deviation from U.S. mean in U.S. SD units
Listening Comprehension	49	-8.0
English Structure	49	-8.1
Vocabulary	48	-4.4
Total	484	-4.0
Reading Comprehension	48	-1.5
Writing	48	-1.1

These differences in degree of overlap constitute evidence of differential levels of development (toward native-speaker performance levels) in the proficiency domains being assessed directly and indirectly by the TOEFL sections. Note that for TOEFL examinees, the TOEFL Total mean was some 4.0 standard deviations below the corresponding U.S. examinee mean. However, it is evident that section means were not equally low when similarly scaled relative to corresponding U.S. distributions. In fact, the section means ranged from -1.1 (Writing) and -1.5 (Reading Comprehension) to approximately -8.0 (English Structure and Listening Comprehension). These findings indicate, among other things, that EFL users/learners who take the TOEFL Test tend to have developed substantially more "native-like" levels of proficiency in the domains assessed by Writing and Reading Comprehension, than in the domains assessed by English Structure and

Listening Comprehension, for example. For present purposes, attention may be focused on the relatively depressed LC mean, from which it may be inferred that for the typical TOEFL examinee, functional ability to comprehend spoken utterances in English is less well developed (from a native-speaker performance perspective) than are those involved in reading—with comprehension—material written in English, for example. And because development of speaking proficiency in a target language is linked functionally to, and will tend to be contingent upon prior development of, the functional ability to understand what is being said in that language, a corollary inference may also be drawn. More specifically, it seems plausible that average latent levels of, say, ILR-scaled speaking proficiency in representative general samples of TOEFL examinees, as well as in samples by native language would not tend to exceed (although they would tend to co-vary strongly with) similarly scaled average levels of listening comprehension. Accordingly, these findings are interpreted as being consistent with those cited above (Carroll, 1967). For a relatively detailed treatment of the findings reported by Carroll (1967), as well as of the interpretive contribution of referring TOEFL scores to native-speaker norms, see Wilson (1989, pp. 11-16). It is important in the general context here under consideration to note that whereas the native-speaker frame of reference permits the inference that distributions of underlying abilities tapped by the TOEFL sections are *not* comparable (consistent with ETS, 1973, p. 11), scores on the TOEFL sections involved nonetheless were moderately highly intercorrelated (e.g., ETS, 1973, p. 15, Table 3). It is also pertinent to note that the highest correlation reported ($r = .78$) was for English Structure versus Writing, two aspects of proficiency with respect to which TOEFL examinees were seen, above, to be markedly different in terms of inferred developmental level. The fact that the skills being tapped by these two sections were found to have been relatively closely related from a correlation perspective suggests that they were taught or learned together. The fact that the learners exhibited markedly different levels of attainment in the skill domains involved when their performance was referenced to that of native speakers *suggests primarily that the skills being tapped by the two sections were not learned at the same rate*.

6. Such a differential pattern could be accounted for by, for example, relatively less curricular emphasis on spoken English in the Korean educational programs than in the EFL instructional programs in other countries, including those represented in the study sample. And it is possible, for example, that instruction tends to be more effective in the area of written

communication than in the area of spoken communication in the Korean environment. It is also possible that, due to the nature and structure of the respective languages, development of proficiency in oral communication in English may tend to be relatively more difficult for native-Korean speakers, on the average, than for native speakers of other languages, including those represented in the TOEIC sample. Explication of the pattern is clearly beyond the scope of this inquiry.

7. Raw scores on the three TOEFL sections are converted to comparable standard scales ranging, in theory, between 20 and 80, with a mean of 50. In practice, the reported scaled-score means for general reference groups tend to center around 50 and range between 20 and 70, with slight variation by section. The Total score is an additive function of the three section scores with a theoretical 200–800 range. In practice, distributions of total scores for TOEFL general reference groups tend to center around 500 and vary between roughly 250 and 677 (see, for example, ETS, 1981, p. 14). For present purposes it is important to keep in mind that, as noted by the TOEFL program (e.g., ETS, 1973, p. 13), ". . . a reported part score of 50 on any form of the TOEFL has no absolute meaning with respect to level of proficiency in English." This, of course, is applicable to any given part (section) or Total score on the TOEFL Test, or any similarly constructed, norm-referenced test of English proficiency, such as the TOEIC Test, for example (see Note 5, above, for further development of this point).

8. What is implied here (and hereafter) is that the observed means tend to reflect characteristic patterns of differences in foreign-language-learning outcomes for typical native speakers of other languages who have initiated curriculum-based formal study of English as a foreign language after reaching linguistic maturity in the mother tongue. According to one recent source (CFLCFCC, 1999, p. 4), "[i]n most European countries and Japan and Korea, five to seven years generally are allocated to the study of another language." It is plausible that English is one of the foreign languages involved in most instances. No comprehensive, systematically developed comparative data on rates of acquisition of EFL proficiency acquisition by native speakers of other languages could be located during the course of this study.

9. Generalizability beyond the "TOEFL examinee" context, for example, to populations of prospective TOEIC test takers, appears to be plausible based on the following lines of reasoning

and evidence. First, although the TOEFL Test differs from the TOEIC Test with respect to item content and item types, standard score scales, and so forth, both tests appear to be measuring generally similar aspects of ESL proficiency, namely, listening comprehension versus reading comprehension skills and knowledge of appropriate English language usage, for example. Both tests include sections made up of items that involve primarily spoken stimuli and sections with items that involve only written stimulus material; the corresponding scores are quite closely related (e.g. ETS, 1985a; Hemingway, 1999; Wilson, 1989; Wilson, Berquist, & Bell, 1998; Woodford, 1982;). Second, it appears to be a plausible assumption (e.g., Wilson, 1989, p. 27) that *within* subpopulations of educated, EFL/ESL test takers defined by native country and/or native language, those who take the TOEIC Test in places of work or work-related ESL training and those who take the TOEFL Test in connection with their undergraduate- or graduate-study plans, tend to have had generally comparable EFL/ESL-learning opportunities. By inference from the foregoing, TOEFL- and TOEIC-takers *within* various national/linguistic contexts presumably tend to have had generally comparable ESL learning backgrounds, and are differentially selected into the respective test-taking populations primarily by variables associated with differences in career orientation. Finally, given the foregoing assumptions, it follows that we would expect representative TOEIC and TOEFL examinee subgroups to exhibit generally similar patterns, although not necessarily similar levels of relative average ESL-proficiency attainment across domains being tapped in somewhat differing ways by the respective tests—e.g., relative attainment in aural versus reading- and writing-related domains at time of testing. We also assume that inferences about language group differences in patterns of relative development of EFL macroskills that are based on evaluation of the performance of TOEFL native-language reference groups on TOEFL sections will tend to be generalizable to such reference groups in the TOEIC testing context. Emphasis on TOEFL means for native-language groups, rather than on TOEIC means for corresponding groups, stems primarily from the fact that much more comprehensive data are available for the TOEFL Test (which has been widely used since the late 1960s) than for the TOEIC Test, which was introduced in Japan in 1979 (see, for example, Woodford, 1982) and designed to be used primarily in intracountry assessments (see CGI, 1996, for a summary of data on TOEIC test performance of examinees classified by selected background variables and country of residence).

10. A comprehensive set of background questions has been included as a standard feature of TOEIC test administrations for several years (e.g., CGI, 1996). Questions about years of study of English as a foreign language and the highest level of education attained by an examinee are included. Results of unpublished internal analyses indicate that each contributes independently to prediction of TOEIC test performance. For reasons suggested above (see Note 9), it was not feasible to use TOEIC-generated native-language reference-group data (including background questions) for the present study.

11. Consider, for example, a recent statement by the CFLFCC (1999, pp. 32-34, passim, emphasis added):

> Language teachers *know* that some languages are more difficult to learn than other languages. [Hence] the same amount of instruction in different languages may not result in similar levels of proficiency *even under ideal conditions of instruction and with motivated learners*. . . . In sum, foreign language teachers and administrators must take these differences in language difficulty into account when planning to offer *more difficult languages* in their foreign language program. In doing so, foreign language educators should develop assessment outcomes that are language specific, and *communicate realistic expectations* to parents, students and administrators.

Such planning presupposes, as a minimum condition, the availability of a valid classification of other languages according to relative learning difficulty when studied as foreign languages by native-English speakers.

12. Because the DLIFLC classification has been directly validated, it is noteworthy in the context of the present study that findings such as those shown illustratively in Table 13 (from Wilson, 1995) incidentally lend added indirect support to working assumptions about the potential usefulness of TOEFL reference-group means as group-analytic variables that tend to convey valid information about language-group-related differences with respect to rates and patterns of EFL learning. More direct assessments such as those suggested herein, for example, are needed.