

DOCUMENT RESUME

ED 093 948

TM 003 766

AUTHOR Heil, Donald K.; Aleamoni, Lawrence M.
TITLE Assessment of the Proficiency in the Use and Understanding of English by Foreign Students as Measured by the Test of English as a Foreign Language.

INSTITUTION Illinois Univ., Urbana. Office of Instructional Resources.

REPORT NO RR-350
PUB DATE Mar 74
NOTE 38p.

EDRS PRICE MF-\$0.75 HC-\$1.85 PLUS POSTAGE
DESCRIPTORS *English (Second Language); *Foreign Students; Graduate Students; Language Ability; *Language Tests; *Predictive Ability (Testing); *Predictive Validity; Test Validity

IDENTIFIERS English Placement Examination (EPE); Test of English as a Foreign Language; TOEFL

ABSTRACT

The grades which foreign students receive are not always based on the same criteria as the grades assigned to native American students. The use of standardized test scores provides a common data base from which to evaluate the relative proficiency level of foreign students. This study examines the Test of English as a Foreign Language (TOEFL) and the English Placement Examination (EPE), the two tests most frequently used to place foreign students. The objectives of this study are as follows: (1) to determine the predictive ability of the TOEFL and the EPE using first and second semester graduate grade point average and also using a grade in a remedial English course as the criterion, (2) to assess the concurrent validity of the TOEFL and the EPE, and (3) to estimate the magnitude of the change in the TOEFL scores after the subjects have lived in an English speaking country and taken a remedial English course for one semester. (BB)

8766607

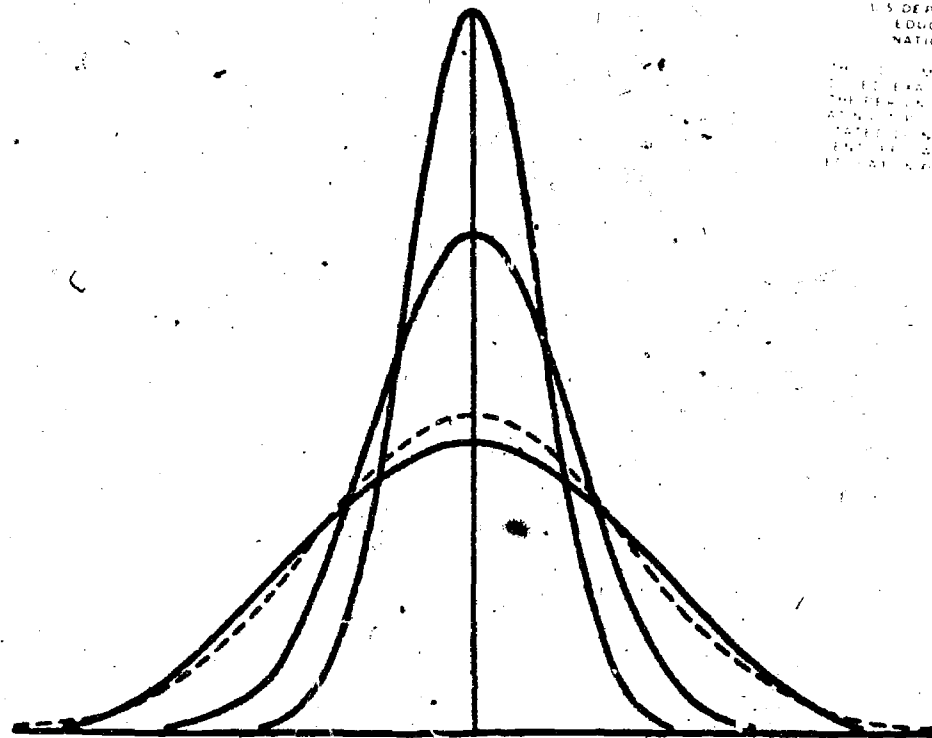
Research Report

TITLE

ASSESSMENT OF THE PROFICIENCY IN THE USE AND UNDERSTANDING
OF ENGLISH BY FOREIGN STUDENTS AS MEASURED
BY THE TEST OF ENGLISH AS A FOREIGN LANGUAGE

AUTHOR(S)

Donald K. Heil and Lawrence M. Alesoni



U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
1201 KENNEDY DRIVE
WASHINGTON, D.C. 20004

REPORT # 350

DATE March 1974

May be quoted in whole or in part if credit is given the source.

MEASUREMENT AND RESEARCH DIVISION - OFFICE OF INSTRUCTIONAL RESOURCES
307 ENGINEERING HALL - UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801

ASSESSMENT OF THE PROFICIENCY IN THE USE AND UNDERSTANDING OF
ENGLISH BY FOREIGN STUDENTS AS MEASURED BY THE
TEST OF ENGLISH AS A FOREIGN LANGUAGE

Most prediction studies of native born students indicate that previous academic record is the best predictor of further college achievement (Bowers, 1965; Duff & Aukes, 1965). However, the academic records of foreign students have been found to be extremely diverse in type and content depending on the country of origin and the institution within the country. Thus, most efforts to use these diverse records as predictors of academic success in American colleges have been unfruitful (Putman, 1961). Since the previous academic records were not available for a large number of students in the present study, no attempt was made to use the remaining available academic records.

Howell (1966) suggested that standardized test scores could provide a common data base from which to evaluate the relative proficiency level of the entering foreign students. This data could then be used for prediction purposes.

A number of investigators have attempted to predict academic success, i.e., grade-point average (GPA) of foreign students from typical American college standardized admission tests. Pitcher and Harvey (1963) studied the predictive use of the Graduate Record Examination (GRE). Coffman (1963) reported the predictive value of the College Board Preliminary Scholastic Aptitude Test (SAT). Howell (1966) studied the predictive value of the scholastic aptitude and achievement tests. Other tests, such as the Miller Analogies Test, have been studied for their predictive capabilities (Houtras, 1956). Sims (1967) summarized the studies in this area and reported a median correlation of about .22 for verbal ability test scores and about .28 for quantitative ability scores. Sims further pointed out that these studies did not assess the foreign student's proficiency in the use and understanding of English. The possible effect of the foreign student's

proficiency in English upon performance on American admission tests appears as a reasonable next step. This line of reasoning would suggest that perhaps the same meaning should not be assigned to a low score on a scholastic achievement or aptitude test obtained by a student with good English proficiency as would be assigned if obtained by a student with low English proficiency. In other words, the proficiency of English may moderate the relationship between scholastic aptitude and achievement.

Slocum (1967) presented the view for the use of the Test of English as a Foreign Language (TOEFL) as a moderator variable. He stated, "... Such tests as the SAT and GRE are most useful in the case of foreign student applicants when their scores are considered in the light of the TOEFL scores. This enables one to make allowances if necessary in objective test scores for a handicap in English. On the other hand, the TOEFL score could point to rejection if it were high (indicating little or no handicaps in English) and the objective test scores (not so much verbal aptitude scores as quantitative and achievement scores) were low [p. 4]."

However, Sharon (1970) reports inconclusive results in a study which tested the hypothesis that TOEFL is a moderator variable when combined with the Admissions Test for Graduate Study in Business (ATGSB) to predict success of foreign students in graduate business school. The TOEFL added nothing to the predictive validity of the ATGSB when combining them in a linear or a nonlinear model. Schrader and Pitcher (1970) combined the Law School Admission Test (LSAT) and TOEFL scores to predict success in law school of foreign students. They reported that the combined use of the LSAT and the TOEFL in a linear regression model did not increase the predictive validity obtained when either test was used alone.

Sharon (1971) tested TOEFL as a moderator of the GRE scores in the prediction of foreign students' GPA in graduate school. This study reports

essentially the same magnitude and configuration of the GRE validities as were found by Pitcher and Harvey (1963) which were .24 between GRE-V and GPA and .32 between GRE-Q and GPA. The more enlightening result reported was that the linear combination of the GRE-V or Q with the TOEFL did not result in significantly higher validities over those obtained with the GRE-V or Q alone. One of the reasons suggested for this finding is that GPA may be an inappropriate criterion for foreign students. For example, in most colleges foreign students with inadequate English proficiency are frequently given special remedial English instruction and, thus, must carry lighter course loads. Sharon further postulates that allowances are often made in the grades themselves to compensate for the inability of the foreign student to deal adequately with the American language, concepts and meaning of the material presented.

The entire question of the factors which contribute to the unreliability of course grades or GPA when they are used as the criteria in prediction studies has been a compounding problem for some time (Ebal, 1961).

Lavin (1965) discusses several other sources of variations in assigning grades. Grades are based on different criteria. Some are based on objective tests, others on essay-type tests. Others include credit for oral participation and term papers. Grades many times include a variety of subjective factors based on undefined criteria. GPAs also have limitations. They vary based on the number of courses used in the computation; the GPA for one student may be based on four courses whereas another student's may be based on six courses. The courses may vary in content, in level and difficulty. The courses most certainly will require differing abilities and levels of proficiency. Lavin further stated that the GPAs of foreign students tend to be based on few courses because as Sharon pointed out, the majority of foreign students carry light course loads because they must enroll in remedial English courses. This is especially true in the early semesters until they meet the English proficiency requirements.

Another factor which may contribute to the unreliability of the grades for foreign students is the "good will" grade. The grades which foreign students receive are not always based on the same criteria as the grades assigned to native American students. Putman (1961) reported that one-third of the faculty of the Teacher's College at Columbia University admitted to some leniency in grading foreign students. The National Association for Foreign Student Affairs has attempted to discourage courtesy grades which are assigned as a gesture of international good will or misplaced compensation for the student's lack of familiarity with the English language (Walton, 1967). Paraskevopoulos and Dremuk (1968) report no differential standards are used for foreign students in the physical and biological sciences, either in course requirement or in grading standard at the University of Illinois at Urbana-Champaign (UIUC). In the social and behavioral sciences where the need for language and communication proficiency is greater, they report that a large minority of the faculty is more lenient in the grading of foreign students from non-English speaking countries and use differential course requirements for foreign students.

These potential contaminants of criteria (grades or GPA) would tend to lower the reliability of the criteria, thus, reducing the correlation between test scores and the criteria (validity).

Most American colleges and universities use the TOEFL scores in conjunction with so-called differential, locally administered, placement tests which, in combination, are used to place the foreign student in remedial English courses. These remedial courses are designed to increase the foreign students' proficiency in English. Studies have been conducted to determine the predictive validity of a number of these tests as related to GPA. Maxwell (1965) investigated the comparative validities of the TOEFL and the English proficiency test used at Berkeley. Using a sample of 238 cases, the correlation of the TOEFL with GPA was .17, the corresponding validity of the Berkeley tests was .11.

The University of Washington (1966) reported similar validity relationships using the TOEFL and the English proficiency test used at Washington. The Spearman rank-order correlation for both tests was .26. A replication was conducted in which an identical rank-order correlation was obtained.

Domino (1966) conducted a validity study at Fresno State College using the TOEFL and the College Vocabulary Test. The validity of the TOEFL vs GPA reported was .31, the corresponding validity of the College Vocabulary Test was .34.

This sample of studies appears to show that the locally administered English proficiency tests produce predictive validities comparable to that of the TOEFL.

In light of this conclusion, the concurrent validity of the TOEFL with the locally administered English placement and proficiency is of interest. Are these locally administered tests merely a redundancy of the TOEFL?

1. Upshur (1966) reports a concurrent validity study using students from Indiana University ($N = 38$), San Francisco State College ($N = 50$) and Park College ($N = 12$) (total $Ss = 100$). He reported a concurrent validity of .89 between the TOEFL and the Michigan Test of English Language Proficiency.
2. Maxwell (1965) reported a study conducted at the University of California, Berkeley campus. He found a concurrent validity correlation of .87 ($N = 238$) between the TOEFL and the locally developed English proficiency test at Berkeley.
3. The American Language Institute (ALI) at Georgetown University (1966) reported a concurrent validity of .79 between the TOEFL score and the ALI test developed at Georgetown.
4. Educational Testing Service (ETS) (1966) conducted a study at New York University comparing the TOEFL with the ALI Test of Proficiency in English developed at New York University. This study also reports the

concurrent validity of the TOEFL with the Michigan Test of English. The correlation of the TOEFL total score with the ALL total score was .81 ($N = 166$) and with the Michigan test, total score was .84 ($N = 99$).

5. Pitcher and Ra (1967) conducted a study comparing the TOEFL scores with ratings of themes written by 210 foreign students enrolled in colleges and universities throughout the United States. Each student wrote four themes, one on each of four assigned topics. Independent ratings of each theme were made by two raters on overall quality. Rater reliability of .92 for the sum of the eight ratings was obtained. A correlation of .78 was reported between the TOEFL total score and sum of the eight ratings. A correlation of .74 was obtained between the Writing Ability subscore on the TOEFL and the sum of the eight ratings.

Most college and university foreign student admission policies require a measure of the student's proficiency in the English language.

Competence in the English language has been assumed to be the crucial factor in the success of the foreign student studying at an American college or university. This view is perhaps based on face validity in that it would be difficult to understand how a foreign student could learn in an American graduate school without adequate proficiency in the English language; to read, write and comprehend the material presented.

Thus, most graduate schools recommend or require that their foreign students take the TOEFL in their native country and submit the results for admission.

Description of TOEFL

The TOEFL consists of five subtests which are: Listening Comprehension, English Structure, Vocabulary, Reading, Comprehension and Writing. A score for each student as well as a total score is provided by ETS.

- 7.
- The ETS manual states the following descriptions of the TOEFL Subtest
- Part 1 Listening Comprehension is a test of the ability to understand spoken English.
 - Part 2 English Structure is a test of recognition and mastery of the important structural and grammatical aspects of spoken English.
 - Part 3 Vocabulary is a test of vocabulary necessary for efficient reading.
 - Part 4 Reading Comprehension is a test of the ability to read and understand English prose; including the ability to make inferences and draw conclusions.
 - Part 5 Writing Ability is a test of effective style, usage and diction in written English.

The TOEFL is scored by ETS; part scores as well as a total score are sent to the examinee and the educational institutions to which the student has applied for admission.

A more specific and detailed description of the TOEFL subtests will illustrate the verbal skills being assessed.

The Listening Comprehension subtest consists of these parts: direct questions, conversation followed by questions, and a lecture followed by questions pertaining to its content.

The second subtest of the TOEFL is English Structure. The 20 items consist of a segment of dialogue containing a blank and four multiple choice alternatives-- the examinee fills the blank with the correct response. This subtest is reported to measure tense, sequence of nouns, adjectives, adverbs, etc.

The third subtest, Vocabulary, consists of two parts, A and B. Part A is of the fill-in-the-blank type with four alternatives provided. Part B consists of synonyms or definitions and four alternatives, one being correct. Part A contains 15 items and Part B, 25 items.

The fourth section of the TOEFL is Reading Comprehension. This subtest consists of a text and questions based on the text. The student may reread the text in order to answer the questions.

The fifth subtest is Writing Ability which has two parts. Part A consists of sentences in which four words or phrases are underlined in each sentence. The underlined words or phrases are labeled A, B, C, and D. The student is to select the one incorrect word or phrase for each sentence. Part B consists of incomplete sentences for which four alternatives are provided. The examinee is to select the correct alternative to complete the sentence. This subtest appears to stress primarily basic English grammar.

The TOEFL was developed in 1963 as a cooperative project of 30 organizations, public and private. A council for the TOEFL was established with members from both private and governmental agencies. The purpose of the council was to direct the testing of English language proficiency of foreign students who apply for admission to United States colleges and universities. In 1965, the administrative functions of the TOEFL council were assigned to the co-sponsorship of the college Entrance Examination Board (CEEB) and ETS.

Objectives

The objectives of the present study are as follows:

1. Determine the predictive validity of the TOEFL using first or second semester graduate GPA as the criterion.
2. Determine the predictive validity of the English Placement Examination (EPE), developed at UIUC, upon the criteria of first and second semester graduate GPA.
3. Assess the concurrent validity of the TOEFL and the EPE.
4. Assess the predictive validity of the TOEFL scores and the EPE scores using grade in a remedial English course as the criterion.

5. Estimate the magnitude of the change in the TOEFL scores after the Ss have lived in an English speaking country and taken a remedial English course for one semester.

GPA was used with the full recognition of the inherent limitations. Lannholm (1967) pointed out the restriction of range in graduate grades. At the graduate level most teachers give only A's and B's. The GPA of foreign students is further contaminated because most take remedial English courses, thus, a reduced load until their proficiency in English has met a prescribed criteria. Even in the light of these limitations, grades are the primary criterion in the assessment of academic achievement; their use in predictive studies is justifiable.

Selection and Placement Procedure for Foreign Students at UIUC

The UIUC requires the foreign student to submit scores for either the English Language Inventory (ELI) or the TOEFL. Those students who obtained a score of 80-89 on the ELI or a score of 480-569 on the TOEFL are required to take the UIUC EPE which is administered by the Division of English as a Second Language (ESL). Those applicants whose scores are above this range are placed in 400-level courses. Those applicants below this range are generally not accepted for admission.

The results of the EPE are used to recommend the appropriate placement in remedial courses of English which were established to supplement the specific deficiency of the foreign student. The placement procedure also is used to advise or control the academic course load the student may carry.

The EPE consists of four parts:

Part 1 Structure is a test of the recognition of English grammar and sentence structure in writing. A score of 100 is perfect; a score of 55 is the lowest score acceptable for the most elementary English course.

Part 2 Aural Comprehension is a test of the understanding of spoken English at normal speed. A score of 100 is perfect; 55 is the lowest acceptable score for the most elementary English course.

Part 3 Original Composition is a 40 minute test of writing on an assigned but familiar topic and outline. The scores range from 0-9; 0 means no problems, 1 means very good, 5 poor, 8 is the lowest acceptable score.

Part 4 Pronunciation is a test of the ability to understand and to be understood orally. The scores range from 1-5; a score of 1 is native sounding, 5 is almost unintelligible.

ESL does not produce a composite or total score as does ETS on the TOEFL. ESL states in its procedures, "the composite scores are not significant because a low score in one test area can reduce total effectiveness; therefore, a student is usually only as good as his lowest test area score." The multiple cutoff method is advised.

Subjects

All Ss were incoming foreign students who had been accepted for admission to UIUC for the fall of 1970.

A total of 154 Ss constituted the initial sample for the September 1970 pretest.

The follow-up sample was reduced to 44 Ss for the February 1971 posttest.

Procedure

In cooperation with the Office of Foreign Admissions, the Measurement and Research Division (MARD) of the Office of Instructional Resources (OIR) administered the TOEFL examination to 154 incoming foreign students in September of 1970.

The Office of Foreign Admissions supplied to MARD the TOEFL scores which the Ss had presented with their application for admission to UIUC. The Ss took this TOEFL in their native countries. The intervening time between

the time of admission and time the Ss took the TOEFL which was submitted with their admission application ranged from two months to six months. The Ss were treated as a whole and intervening time was not controlled.

The Ss in the study then proceeded with the usual procedure for placement in English courses. Six Ss were dropped from the sample because they were undergraduates. One hundred and ten of the remaining 148 Ss had scored below 569 on the TOEFL submitted with their application for admission which made it necessary for them to take the EPE examination.

Table 1 presents the score norms and class recommendations for the four parts of the EPE test. The course load scale is as follows: a student placed in Rhetoric 110 or 110 + 400 or 401 can carry a full academic load. A student placed in Rhetoric 110 + 111 can carry one-half course load, a student placed in Rhetoric 109 + 110 can carry only one-fourth course load. The significance and purpose of the rhetoric courses recommended is set forth in Table 1.

The following distribution of students were assigned to and completed the prescribed remedial English course based on the scores they obtained on the EPE examination.

<u>English Course</u>	<u>No. of Ss</u>
109	2
110	42
111	29
114	1
115	1
400	12
401	2
	<hr/> 89

The remaining 21 students of the 110 students either dropped the English course to which they were assigned or no placement was made for them by ESL.

The Ss' scores on the EPE examination were obtained from ESL.

The next step was to assess the impact of living and functioning in an English speaking environment combined with enrollment in a prescribed remedial English course, on the observed proficiency in English as measured by the TOEFL.

TABLE 1

Grade Norms and Class-Recommendation Significance

Rhetoric Courses Recommended	Structure Score	Aural Comprehension Score	Composition Rating (Writing)	Pronunciation Rating	Significance and Purpose of Courses Recommended
400 ^a	90-78	94-80	3-4		Orientation to basic techniques of graduate study and effective expression. (3 hours per week)
401 ^a	98-90	98-94	1-2		Orientation to advanced problems of graduate study and thesis preparation. (3 hours per week)
111 ^b	83-68	85-68	4-6		Rapid and intensive review of basic structure and study of more complicated sentence patterns, with oral drill and written compositions. (6 hours per week)
109 ^b	72-55	72-55	6-8		Intensive course in basic English structure with oral and written practice. (6 hours per week)
110 ^c		85-55		3-5	Pronunciation. Required if aural comprehension score is 70 or below, or if pronunciation rating is 4 or 5. (3 hours per week plus 2 hours in Language Lab.)

Note.—Students with scores below 55 on structure or aural comprehension or with a composition rating below 8 cannot participate effectively and, therefore, cannot be accepted in courses in ESL at UIUC. Students with a score of 60 or below cannot expect to carry courses other than English.

^aOptional at adviser's discretion.

^bRequired by Graduate College. Expect major language problems.

^cSupplements other courses when needed.

examination. To implement this step, the Office of Foreign Admissions sent a letter to each of the 154 Ss who had taken the TOEFL examination in September 1970 requesting that they participate in the research project and retake the TOEFL examination in February 1971. The students were given a choice of two administration dates in an effort to obtain maximum participation. The response was very disappointing; only 30 out of the original sample agreed to participate, an additional 14 Ss agreed to participate when contacted by a personal telephone request from MARD. Thus, the posttest TOEFL was administered to 44 Ss in February 1971.

The remedial English (Rhetoric 110, 111, 400) course grades and first and second semester GPAs were obtained for each subject.

The validity coefficients were determined for the TOEFL and EPE scores relative to the grade in the rhetoric course and first semester GPA.

Forty-four Ss in the September 1970 TOEFL sample did not submit TOEFL scores with their applications for admission. Thus, the number of Ss used for comparison of the pre-September 1970 and September 1970 TOEFL was reduced to 110.

Results and Discussion

The intercorrelations, means and standard deviations of the TOEFL part scores of the initial sample are presented in Table 2. Variables 1-6 are the TOEFL scores which the Ss submitted with their applications for admissions. Variables 7-12 are the TOEFL scores the Ss obtained in the September 1970 administration.

As can be seen in Table 2, the following are the test-retest correlations (reliabilities) for the part scores and total score of the TOEFL.

TABLE 2
TOEFL Subtest Intercorrelations, Means, and Standard Deviations^a

N = 110

Variables	No.	1	2	3	4	5	6	7	8	9	10	11	Mean	S.D.
Listening Comprehension	1												52.00	10.38
English Structure	2	.372											52.98	6.77
Vocabulary	3	.248	.541										54.90	8.26
Reading	4	.270	.335	.461									55.55	5.60
Writing	5	.378	.689	.646	.498								53.85	7.38
Total	6	.589	.805	.772	.672	.858							539.54	57.21
Listening Comprehension	7	.573	.521	.437	.589	.553	.737						59.98	9.24
English Structure	8	.444	.731	.420	.378	.677	.710	.584					53.82	6.41
Vocabulary	9	.258	.386	.758	.470	.524	.641	.480	.475				53.62	9.01
Reading	10	.472	.397	.417	.571	.454	.622	.706	.496	.486			53.87	7.31
Writing	11	.392	.569	.553	.535	.702	.735	.641	.741	.657	.558		52.38	7.41
Total	12	.524	.625	.645	.630	.707	.845	.847	.784	.775	.795	.874	547.35	64.25

Note.—Correlations of .16 or higher are significant at the .05 level.

^aVariables 1-6 are the Pre-September 1970 TOEFL subtest and Total scores. Variables 7-12 are the September 1970 TOEFL subtest and Total scores.

	<u>r</u>	<u>Number of Items</u>
Part 1 Listening Comprehension	.573	50
Part 2 English Structure	.731	40
Part 3 Vocabulary	.758	40
Part 4 Reading Comprehension	.571	30
Part 5 Writing Ability	.702	40
Total Score	.845	--

In interpreting this table, the reader should bear in mind that the test-retest reliability coefficient is essentially the stability of the *S*'s performance rather than the reliability of the test.

The differences between the means of the TOEFL part scores are presented in Table 3. There was an increase of nearly eight points in the Listening Comprehension ($p < .05$). A significant difference ($p < .05$) was also found in the Total Score. Since we have no knowledge of the student's activities during the intervening time, it is difficult to account for the increase.

TABLE 3

Differences of the Means for the Two TOEFL Administrations

Variables	Pre-September 1970 Mean	September 1970 Mean	Differences	t
Listening Comprehension	52.00	59.98	7.98	9.0 ($p < .05$)
English Structure	52.98	53.82	0.84	NS
Vocabulary	54.90	53.62	-1.28	NS
Reading	55.50	53.87	-1.68	NS
Writing	53.85	52.38	-1.47	NS
Total	539.54	547.35	7.81	9.9 ($p < .05$)

The predictive validities (product moment correlations) of the two administrations of the TOEFL (Pre-September 1970 and September 1970) and the EPE (September 1970) relative to first and second semester GPA criterion measure are presented in Table 4. The TOEFL correlations except for vocabulary on the September 1970 test were significant beyond the .05 level. On the other hand, none of the EPE correlations was significant.

TABLE 4

*TOEFL and EPE Subtest Correlations with First and Second Semester GPAs**N = 110*

No.	Subtest	First Semester GPA	Second Semester GPA
Pre-September 1970 TOEFL			
1	Listening Comprehension	.273	.303
2	English Structure	.347	.319
3	Vocabulary	.183	.279
4	Reading	.179	.226
5	Writing	.390	.340
6	Total	.355	.381
September 1970 TOEFL			
1	Listening Comprehension	.274	.303
2	English Structure	.212	.194
3	Vocabulary	.159	.186
4	Reading	.202	.384
5	Writing	.244	.298
6	Total	.270	.336
EPE September 1970			
1	Structure	.034	.039
2	Aural Comprehension	.068	.123
3	Composition	-.041	-.041
4	Pronunciation	.008	.017

Note.-Correlation coefficients of .16 or higher are significant at the .05 level.

In general, the correlations on the pre-September measures were higher but not significantly higher than those of the September measures for the TOEFL.

It is of interest that the validities of the TOEFL Total scores are higher (nonsignificant) with second semester GPA than with the first semester GPA. The larger number of courses used to compute the second semester GPA may be a contributing factor.

Table 5 shows the predictive validities of the September 1970 TOEFL and EPE part scores with grade in remedial rhetoric courses 110, 111, and 400. For Rhetoric 110 and 111 the grades were Satisfactory Good, coded 4; Satisfactory Fair, coded 3;

Satisfactory, coded 2; Satisfactory Poor, coded 1; and Unsatisfactory, coded 0. In Rhetoric 400, the grades are conventional A through E for credit and were coded A = 5 through E = 0. Only one unsatisfactory grade was given in Rhetoric 110 and none in Rhetoric 111. None of the EPE subscores correlate significantly with grades in Rhetoric 110 or Rhetoric 400 nor is the TOEFL Total score significantly related to grades in the rhetoric courses. The Rhetoric 110 course grade is not significantly correlated with first semester GPA but is significantly related with the second semester GPA. The opposite is shown for Rhetoric 111; whereas the rhetoric course grade for Rhetoric 400 is significantly correlated with both semester GPAs.

TABLE 5

TOEFL, EPE, First and Second Semester GPAs

With Rhetoric Course Grade

No.	Subtest	Rhetoric Course Grade		
		110 ^a	111 ^b	400 ^c
EPE Scores September 1970				
1	Structure	.026	.179	.259
2	Aural Comprehension	.089	.065	-.347
3	Composition	-.100	-.498*	-.417
4	Pronunciation	-.228	.195	.043
TOEFL Scores September 1970				
1	Listening Comprehension	.079	.191	-.459
2	English Structure	.303*	.331*	-.107
3	Vocabulary	-.225	-.060	.054
4	Reading Comprehension	-.108	.211	-.109
5	Writing	-.503*	.174	-.326
6	Total	-.026	.231	-.283
Rhetoric Course Grade				
	First Semester GPA	.257	.352*	.667*
	Second Semester GPA	.345*	.271	.679*

^aN = 40

^bN = 29

^cN = 12

*p < .05

In Rhetoric 110 the TOEFL Writing subscore has a correlation of .503 with the course grade. The EPE scores have lower correlations. The Aural Comprehension correlation was .089. This is to be expected because the students placed in this course were those with low Aural Comprehension scores. The Pronunciation subscore of the EPE had the highest correlation, .228 with grade in Rhetoric 110. Pronunciation improvement is the expressed behavioral objective of Rhetoric 110.

Rhetoric 111 shows a correlation of .498 with the Composition subscore of the EPE. The behavioral objective of Rhetoric 111 is the improvement of written compositions.

The restriction of range of the EPE Composition and Pronunciation subscales should be taken into account in evaluating their correlations.

Table 6 presents the multiple correlations to predict course grade in Rhetoric 110 and 111 and first semester GPA.

*Multiple R to Predict Grade in Rhetoric and First Semester
GPA using EPE Scores and/or TOEFL Part Scores*

Course No.	Criterion Grade in Rhetoric Course Predictors			Criterion First Semester GPA Predictors
	EPE	TOEFL	EPE & TOEFL	EPE & TOEFL
110 ^a	.235	.379*	.703*	.540*
111 ^b	.573*	.467*	.671*	.616*

Note.—The TOEFL total score was *not* used.

^aN = 40

^bN = 29

*p < .05

The EPE (four scores) and the TOEFL (five scores) were used separately and combined as predictor variables to predict course grade in Rhetoric 110 and 111.

Rhetoric 400 was not included because of the low N. For Rhetoric 110, the TOEFL subscores produced the larger multiple R of .379 or compared to .235 for the EPE.

The TOEFL total score was not included in the subscores used in the computation of the multiple R s. All of the multiple R s are significant ($p < .05$) except the EPE scores in Rhetoric 110.

The increase in the multiple R to .703 for the combined TOEFL and EPE in predicting the Rhetoric 110 course grade is significant ($p < .05$) over the multiple R .379 produced when TOEFL was used alone. The increase in the multiple R to .671 using the TOEFL and EPE for Rhetoric 111 is not significant over the multiple R .573 using the EPE subscores alone. Thus, indicating that even when using nine subscores (TOEFL + EPE) to predict Rhetoric 110 course grade, 49% of the variance can be accounted for. Rhetoric 111 presents a somewhat similar picture accounting for 45% of the variance when both TOEFL and EPE's nine subscores are used in computing the multiple R .

The multiple R s obtained using the nine subscores of the TOEFL plus EPE to predict the first semester GPA of the Ss in Rhetoric 110 and Rhetoric 111 were significant ($p < .05$), although accounting for only 29% of the variance in Rhetoric 110 and 38% of the variance in Rhetoric 111.

Tables 7 and 8 present a summary of the stepwise multiple regression of EPE and TOEFL part scores (separately and combined) on grades in Rhetoric 110 and 111, respectively. In Tables 7 and 8, the increase in R^2 is negligible after the third step when using the nine scores of TOEFL and EPE combined to predict course grade.

The three variables entered for Rhetoric 110 are (TOEFL) Writing, (TOEFL) English Structure and (EPE) Composition, in that order. The three variables entered for Rhetoric 111 are Reading Comprehension, Aural Comprehension, and Composition, in that order.

Table 9 presents the test-retest reliability correlations for the 44 Ss that took the posttest February 1971 TOEFL and the September 1970 TOEFL. The test-retest correlations are as follows:

TABLE 7

*Summary Table for Stepwise Multiple Regression of
EPE and TOEFL Scores on Grade in Rhetoric 110*

N = 40

No.	Variable Entered	Multiple		Increase in
		R	R ²	R ²
EPE Only				
1	Pronunciation	.228	.052	.052
2	Composition	.233	.054	.002
3	Aural Comprehension	.235	.055	.001
4	Structure	.235	.055	.000
TOEFL Only				
1	Listening Comprehension	.280	.078	.078
2	Reading Comprehension	.374	.140	.062
3	English Structure	.378	.143	.003
4	Writing	.379	.144	.001
5	Vocabulary	.379	.144	.000
TOEFL and EPE				
1	TOEFL Writing	.502	.252	.252
2	TOEFL English Structure	.601	.361	.109
3	EPE Composition	.675	.455	.094
4	TOEFL Listening Comprehension	.687	.472	.017
5	EPE Structure	.695	.483	.011
6	EPE Pronunciation	.700	.490	.010
7	TOEFL Reading Comprehension	.703	.494	.004
8	EPE Aural Comprehension	.703	.494	.000
9	TOEFL Vocabulary	.703	.494	.000

TABLE 8

Summary Table for Stepwise Multiple Regression of EPE and TOEFL Part Scores (Separately and Combined) on Grade in Rhetoric 111

N = 29

No.	Variable Entered	Multiple		Increase in R^2
		R	R^2	
EPE Only				
1	Composition	.498	.238	.238
2	Structure	.543	.295	.057
3	Pronunciation	.566	.320	.025
4	Aural Comprehension	.573	.328	.008
TOEFL Only				
1	English Structure	.331	.110	.110
2	Listening Comprehension	.409	.167	.057
3	Vocabulary	.465	.216	.049
4	Reading Comprehension	.467	.218	.002
5	Writing	.467	.218	.000
TOEFL and EPE				
1	TOEFL Reading Comprehension	.497	.249	.249
2	EPE Aural Comprehension	.587	.344	.095
3	EPE Composition	.639	.408	.064
4	TOEFL Writing	.650	.423	.015
5	EPE Structure	.656	.430	.007
6	TOEFL Vocabulary	.662	.438	.008
7	TOEFL Listening Comprehension	.668	.446	.008
8	TOEFL English Structure	.669	.448	.002
9	EPE Pronunciation	.671	.450	.002

TABLE 9
 Interrelation and Test Statistics of the TOEFL Test Scores for the September 1970
 Administration (Pretest) and TOEFL Scores for the February 1971 (Posttest)
 Administration and the EPE Scores for the September 1970 Administration
 N = 44

Variable	September 1970 TOEFL															Mean	S.D.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Listening Comprehension																59.84	8.52
English Structure	.599															53.25	6.29
Vocabulary	.504	.537														53.77	8.98
Reading Comprehension	.683	.415	.470													54.32	7.14
Writing	.685	.781	.711	.454												50.34	7.16
Total	.854	.796	.806	.740	.884											543.05	62.27
February 1971 IOEFL																	
Listening Comprehension																63.41	8.15
English Structure	.774	.477	.359	.499	.520	.646										57.57	5.90
Vocabulary	.491	.595	.285	.251	.515	.513	.527									54.64	9.80
Reading Comprehension	.593	.626	.730	.478	.723	.775	.595	.538								55.52	6.05
Writing Ability	.572	.243	.384	.606	.321	.530	.511	.371	.642							53.02	6.61
Total	.687	.826	.631	.525	.763	.833	.614	.680	.795	.428						568.31	59.89
	.770	.684	.611	.580	.714	.822	.810	.743	.901	.719	.869						
September 1970 EPE																	
Structure																75.97	14.79
Aural Comprehension	.219	.684	.031	-.004	.442	.324	.283	.653	.391	.024	.567	.462				66.91	11.84
Composition	.505	.545	.246	.156	.438	.482	.677	.597	.509	.142	.703	.649	.595			4.41	1.46
Pronunciation	-.425	-.379	-.392	-.160	-.421	-.465	-.546	-.598	-.602	-.340	-.507	-.661	-.586	-.556		2.94	0.88
	-.500	-.275	-.350	-.232	-.297	-.441	-.459	-.368	-.310	-.247	-.407	-.441	-.294	-.472	-.466		

Part 1	Listening Comprehension	.774
Part 2	English Structure	.595
Part 3	Vocabulary	.730
Part 4	Reading Comprehension	.606
Part 5	Writing Ability	.763
Total Score		.822

These test-retest reliabilities are similar to those obtained in Table 2 for the 110 Ss.

The differences between the means of the TOEFL part scores are presented in Table 10. There was an increase in all of the means. Three increases were significant ($p < .05$), Listening Comprehension, English Structure and Writing. Total was also significant. The problem of assessing the meaning of the gain scores will not be attempted in this paper. As is well known, there are numerous problems in interpreting gain scores (Cronbach and Furley, 1970; Lord, 1956, 1958). Perhaps it would seem reasonable to expect some increase in the scores if the TOEFL does measure proficiency in the use of English just from the fact that the subjects have been functioning in an English speaking environment from September 1970 until February 1971. Table 11 presents the same correlation for the 44's as is presented in Table 4 for the 110's. The mean for first semester was 4.51 and for second semester was 4.54 while the standard deviation for first semester was .461 and .454 for the second semester. The correlation of first and second semester GPA was .803.

TABLE 10

Differences of Means for the Pre and Post TOEFL Administrations

N = 44

Variable	Pretest ^a Mean	Posttest ^b Mean	Differences	T
Listening Comprehension	59.84	63.41	3.57	4.0 (p < .05)
English Structure	53.25	57.57	4.32	5.3 (p < .05)
Vocabulary	53.77	54.64	0.87	NS
Reading Comprehension	54.32	55.52	1.20	NS
Writing Ability	50.34	53.02	2.68	3.0 (p < .05)
Total	543.05	568.31	25.26	2.6 (p < .05)

^a Pre = September 1970^b Post = February 1971

TABLE 11

Correlations of Pre and Post TOEFL Scores, EPE Scores with First and Second Semester GPAs

Subtest	First Semester GPA	Second Semester GPA
Pretest TOEFL		
Listening Comprehension	.225	.303
English Structure	.248	.329
Vocabulary	.096	.090
Reading Comprehension	.076	.195
Writing Ability	.380	.423
Total	.253	.321
Posttest TOEFL		
Listening Comprehension	.199	.262
English Structure	.301	.210
Vocabulary	.219	.217
Reading Comprehension	.057	.061
Writing Ability	.258	.325
Total	.259	.275
September 1970 EPE		
Structure	.096	.166
Aural Comprehension	.236	.346
Composition	-.074	-.091
Pronunciation	.224	.223

In interpreting these tables (10 and 4) the reader should bear in mind the possible motivational problem, since motivation is an important factor.

In both the September 1970 and February 1971 administrations, the students were aware they were only taking the test as part of a research study. Therefore, motivation or the lack of it is a factor to be considered. The test administrators hoped that the desire to do well would minimize the effect of the lack of interest or reinforcement value of the scores. Furthermore, the TOEFL is a long test requiring 140 minutes so fatigue is a probable factor. Finally, since the interval between the test and the retest was long (five months) error of measurement should not be confused with real changes in the students' ability as a result of learning.

The intercorrelation of the TOEFL and the EPE subscores obtained in September 1970 are presented in Table 12.

As might be expected, the EPE Structure subscore correlates .591 with the English Structure subscore of the TOEFL. The EPE Aural Comprehension subscore also correlates .510 with the EPE Structure subscore. The intercorrelations of the four EPE subscores are all significant ($p < .05$). The intercorrelations of the five TOEFL subscores are also all significant ($p < .05$).

Table 13 presents the intercorrelations and test statistics of the TOEFL test parts administered September 1970 for the original sample of 154 subjects. Neither Table 9 nor 10 agree with the ETS manual (1970) which states the "Listening Comprehension is measuring some aspect of English proficiency different from that measured by the other four parts, since the correlations of the Listening Comprehension with each of the others are the lowest coefficients." ETS maintains that "each of the parts contributes something unique to the total, however, is evidenced by the fact that none of the correlations between the part scores is as high as the reliabilities of the part scores."

TABLE 12

Correlations of the EPE Part Scores and the TOEFL
Part Scores September 1970 Administrations

$N = 110$

Subtest	Variable Number	1	2	3	4	5	6	7	8	9	10
EPE Part Scores											
Structure	1										
Aural Comprehension	2	.510									
Composition	3	-.336	-.414								
Pronunciation	4	-.218	-.410	-.421							
TOEFL Part Scores											
Listening Comprehension	5	.156	.515	-.408	-.412						
English Structure	6	.591	.368	-.243	-.183	.332					
Vocabulary	7	.044	.094	-.372	-.120	.281	.304				
Reading	8	.129	.290	-.204	-.214	.510	.179	.276			
Writing	9	.467	.373	-.364	-.227	.396	.584	.561	.274		
Total	10	.336	.470	-.474	-.345	.745	.615	.699	.654	.778	

Note.—The correlation of .16 or higher is significant at the .05 level.

TABLE 13

*Intercorrelations and Test Statistics of the TOEFL
Test Part Scores for the September 1970 Administration*

$N = 154$

Subtest	Variable Number	1	2	3	4	5
Listening Comprehension	1					
English Structure	2	.396				
Vocabulary	3	.453	.520			
Reading Comprehension	4	.579	.447	.488		
Writing Ability	5	.549	.744	.692	.532	
Total	6	.725	.781	.798	.768	.877

To summarize, the predictive validities relative to GPA (product-moment correlations) of the two administrations of the TOEFL (Pre-September 1970 and September 1970) except for Vocabulary on the September 1970 test were significant beyond the .05 level. Whereas none of the EPE subscore correlations with GPA were significant.

The predictive validities of the TOEFL Total scores were higher (nonsignificantly) with second semester GPA than with the first semester GPA. The possible reason is the increased number of courses used to compute the second semester GPA.

None of the EPE subscore correlations with grades in Rhetoric 110 or 400 were significant. Furthermore, the TOEFL Total score correlations with grades in the Rhetoric course were not significantly related.

The Rhetoric 110 course grade is not significantly related with the first semester GPA but is significantly correlated with the second semester GPA. The relationship for Rhetoric 111 is reversed, but the Rhetoric 400 course grade is significantly correlated to the GPA of both first and second semester.

All of the multiple R_s were significant ($p < .05$) except the EPE scores in Rhetoric 110. The increase in the multiple R to .703 for the combined TOEFL and EPE in predicting the Rhetoric 110 course grade is significant ($p < .05$) over the multiple R produced when TOEFL is used alone. This would suggest that the EPE scores are contributing some unique variance not present in the TOEFL scores. The multiple R_s to predict first semester GPA using the EPE and TOEFL are significant for those students in Rhetoric 110 and 111.

Summary and Conclusions

The predictive validities of the TOEFL appear to be similar to the predictive validities for native American students using admissions test scores as reported by Bowers (1965). This is not to suggest that we should expect similar predictive validity levels using different predictors but only that it seems very optimistic to expect to be able to predict academic success for foreign students at a higher level than we are able to predict for native American students. Chase (1972) suggests that the correlations of the TOEFL test parts with the criteria of grades or GPA might be improved if the subtests measured skills that were more highly related to those skills that the student would use in his college work.

To illustrate, let us discuss the probable skills tested in each subtest of the TOEFL separately. The Listening Comprehension Test, parts 1 and 2, are general in content and only slightly tap the verbal skills that measure the student's ability to function in the American culture. Furthermore, the first part concentrates on personal responses by using 16 American given names in 20 items. This type of response has little generality to the skills of the student. The third part of this subtest, which is a lecture, more adequately taps the skills used by the students. If the intention was to measure the understanding of English, Chase would suggest that this test part is measuring the student's ability to take notes and recall details, general intelligence, or the student's previous knowledge of the topic of the lecture and not the understanding of spoken English. However, the understanding of spoken English for the student is an important skill.

The English Structure subtest deals with tense, sequence of nouns and adjectives; the rationale for the selection of the Structure included in the test seems obscure. The formal rules of grammar are clearly important in communication, however, we are all aware that much communication is carried on without loss of content or meaning even when these formal rules are violated. Chase concludes, that the language behavior required by this subtest does not simulate behavior required by the foreign student in an American college.

Part 3, the Vocabulary subtest, is a rather meager sample of the total vocabulary needed by the student. Furthermore, this part correlates .72 with the English Structure subtest. This is not surprising considering the obvious overlap in format of both subtests. Chase suggests that a better format could be developed for this subtest, possibly selecting synonyms which would provide a more adequate sampling of vocabulary, without the burden of sentence reading. Chase further points out that there are many word counts available which could be used to produce a more useful and meaningful test by selecting words based on graded difficulty and abstractness.

The fourth part, Reading Comprehension, includes a short text and questions based on the text. In this part of the test, Chase concludes that the skills required for this test, namely to scan the questions and skim the text for the answers, are important to the student in college work.

The fifth subtest, Writing Ability, is made up of two parts. The first, Part A, contains sentences with four words or phrases underlined in each sentence. The student is to select the incorrect underlined word or phrase. Part B is made up of incomplete sentences with four options presented for the completion of each sentence. Unfortunately, this part of the test is again bound by the formal rules of basic grammar, which does not really indicate the student's understanding of the ideas and concepts conveyed. Furthermore, this subtest correlates in the .70's with several of the other subtests. Chase concludes again that this test does not measure a skill which the student will use in his college activities. Furthermore, the ability to recognize inconsistency in grammatical forms does not really indicate how the student will perform in his own writing. Whereas the student's ability to write and convey ideas and concepts will be used by college instructors to determine in part his course grade. The ability to write effectively is probably reflected to some extent in most of the course requirement and thus, course grades. Chase suggests that a more adequate writing sample format would reflect more closely the writing skills required of the student in his course work in college.

The following is highly speculative and is based on a limited sample of anecdotal input. Foreign students appear to draw a distinction between English and American English. They convey that their difficulty with the TOEFL was that it measures American English, formal grammar sentence structure, and deals little with the ability to understand and communicate in the language. This is not to say that formal grammar is not important in communication but, as was mentioned earlier, communication can be carried on without absolute adherence to the formal rules of American grammar.

Furthermore, some foreign students state that they feel the TOEFL is culturally bound by using formats such as the personal responses of the first part of the Listening Comprehension subtest. This may be a valid criticism of the TOEFL and may be limiting the correlation with GPA or grades. Thus, indicating the foreign student can perform successfully in an American college without strict adherence to formal rules of grammar and structure of American English.

In conclusion, the TOEFL appears to be no better or worse than those admission tests used to predict success for native American students. Perhaps it could be improved by measuring more of the skills the foreign student will be required to use in his daily academic performance.

References

- American Language Institute. A report of the results of English testing during the 1966 pre-university workshop at the American Language Institute. Unpublished manuscript, Georgetown University, 1966.
- Bowers, J. A five-semester follow-up of fall 1962 beginning freshmen at the University of Illinois. Research Memorandum 65-4. Urbana, Illinois: University of Illinois, Office of the University Dean of Admissions and Records, 1965.
- Chase, C. I. Review of Buros (Ed.), *Test of English as a Foreign Language. The Seventh Mental Measurements Yearbook*, 1972, 1, 266-267.
- Coffman, W. E. Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude. Princeton, New Jersey: Educational Testing Service for CEEB, June 1963.
- Cronbach, L. J. & Furley, L. How we should measure "change" - or should we? *Psychological Bulletin*, 1970, 74(1), 68-80.
- Domino, G. Academic achievement and English proficiency of foreign students as predicted by the Test of English as a Foreign Language. Unpublished manuscript, Fordham University, 1966.
- Duff, F. L. & Aukes, L. E. The relationship of the Graduate Record Examination to success in the graduate college. Research Memorandum No. 103. Urbana, Illinois: University of Illinois, Bureau of Institutional Resources, 1965.
- Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1961, 16, 407-424.
- Educational Testing Service. *Test of English as a Foreign Language: Interpretive Information*. Princeton, New Jersey, 1970.

- Educational Testing Service. Test of English as a Foreign Language (TOEFL) Research - correlations among TOEFL, American Language Institute and Michigan tests. Unpublished manuscript, Princeton, New Jersey, 1966.
- Houstras, P. T. Improving our selection of foreign graduate students. *The North Central Association Quarterly*, October 1956.
- Howell, J. J. Prediction of freshmen grade point average of foreign students at U. S. institutions of higher learning--a pilot study. Pennington, New Jersey: Prepared for CEEB by New Jersey Educational Research, April 1966.
- Lannholm, G. V. Graduate schools requiring or recommending GRE scores for initial admission to graduate study. Graduate Record Examination Special Report 67-3. Princeton, New Jersey: ETS, 1967.
- Lavin, D. E. The prediction of academic performance. New York: Russell Sage Foundation, 1965.
- Lord, F. M. The measurement of growth. *Educational and Psychological Measurement*, 1956, 16, 421-437.
- Lord, F. M. Further problem in the measurement of growth. *Educational and Psychological Measurement*, 1958, 18, 437-454.
- Maxwell, A. A comparison of two English foreign language tests. Unpublished manuscript, University of California (Davis), 1965.
- Paraskevopoulos, J. & Dremuk, R. Grading patterns for foreign students: A faculty survey. Research Memorandum 68-2. Urbana, Illinois: University of Illinois, Office of the University, Dean of Admissions and Records, 1968.
- Pitcher, B. & Harvey, P. The relationship of the GRE Aptitude Test scores and graduate school performance of foreign students at four American graduate schools. Special Report 63-1. Princeton, New Jersey: Graduate Record Examinations, ETS, April 1963.

- Pitcher, B. & Ra, J. B. The relation between scores on the Test of English as a Foreign Language and ratings of actual theme writing. Statistical Report 67-9. Princeton, New Jersey: ETS, 1967.
- Putman, I., Jr. The academic performance of foreign students. *The annals of the American Academy of Political and Social Science*. May 1961.
- Schrader, W. B. & Pitcher, B. Interpreting performance of foreign law students on the Law School Admission Test and the Test of English as a Foreign Lanugage. Statistical Report 20-25. Princeton, New Jersey: ETS, 1970.
- Sharon, A. T. The effectiveness of TOEFL and ATGSB as combined predictors of foreign student success in graduate business schools. Unpublished manuscript. Princeton, New Jersey: ETS, 1970.
- Sharon, A. T. Test of English as a Foreign Language as a moderator of Graduate Record Examination scores in the prediction of foreign students' grades in graduate school. R 13-71-50. Princeton, New Jersey: ETS, 1971.
- Sims, A. G. Research related to the admissions of foreign students to U. S. colleges and universities. Paper presented at the Seminar on Research in Programs for Foreign Students, Michigan State University, Lansing, 1967.
- Slocum, J. B. Using TOEFL in foreign student admissions. Paper presented at the annual meeting of CEEB, 1967.
- University of Washington. Foreign student study. Unpublished manuscript, 1966.
- Upshur, J. A. Comparison of performance on Test of English as a Foreign Language and Michigan Test of English Language Proficiency. Unpublished manuscript. University of Michigan, 1966.
- Walton, B. J. Foreign student research and its implications. Washington, D. C.: U. S. Department of State, Office of External Research, 1967, Part III, p. 26.