DOCUMENT RESUME

ED 205 607 TM 810 517

AUTHOR Angelis, Paul J.; And Others
TITLE The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-79-7; TOEFL-RR-3
PUB DATE Oct 79
NOTE 53p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Aptitude Tests; *College Admission; *Comparative Analysis; English (Second Language); Higher Education; *Native Speakers; *Non English Speaking; *Verbal Ability
IDENTIFIERS Graduate Record Examinations; Scholastic Aptitude Test; *Test of English as a Foreign Language; Test of Standard Written English

ABSTRACT

The performance of two groups of non-native English speakers on the Test of English as a Foreign Language (TOEFL) and an appropriate verbal aptitude test was examined. One group of graduate applicants took both TOEFL and the verbal section of the Aptitude Test of the Graduate Record Examinations (GRE). Another group of undergraduate applicants took TOEFL, the verbal section of the College Board Scholastic Aptitude Test (SAT), and the Test of Standard Written English (TSWE). Data are presented showing how native and non-native speakers compare on each set of tests. Information is also provided to aid in interpreting test results for non-native speakers who have taken both types of tests. The appendix to the report summarizes item reviews, by specialists in English as a Second Language, which suggest future directions for TOEFL test development. (Author/GK)

TOEFL Research Reports

REPORT 3
OCTOBER 1979

THE PERFORMANCE OF NON-NATIVE SPEAKERS OF ENGLISH ON TOEFL AND VERBAL APTITUDE TESTS

Paul J. Angelis
Spencer S. Swinton
William R. Cowell

EDUCATIONAL TESTING SERVICE

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and, in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; Graduate Record Examinations Board members are associated with graduate education.

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

The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests

Paul J. Angelis

Spencer S. Swinton

William R. Cowell

Educational Testing Service
Princeton, N.J.

RR 79-7

Abstract

This study examined the performance of two groups of non-native English speakers on the Test of English as a Foreign Language (TOEFL) and an appropriate verbal aptitude test. One group of graduate applicants took both TOEFL and the verbal section of the Aptitude Test of the Graduate Record Examinations (GRE). Another group of undergraduate applicants took TOEFL, the verbal section of the College Board Scholastic Aptitude Test (SAT), and the Test of Standard Written English (TSWE). Data are presented showing how native and non-native speakers compare on each set of tests. Information is also provided to aid in interpreting test results for non-native speakers who have taken both types of test. The appendix to the report summarizes item reviews, by specialists in English as a Second Language, which suggest future directions for TOEFL test development.

Table of Contents

The Performance of Non-native Speakers of English on TOEFL and Verbal Aptitude Tests

Previous studies (e.g., Angoff & Sharon, 1970; Clark, 1977) have shown that the Test of English as a Foreign Language (TOEFL) does clearly distinguish between native and non-native speakers of the language. Native speakers perform exceedingly well on TOEFL, finding little difficulty with any section of the test. Non-native speakers, on the contrary, consistently show varying achievement on TOEFL, their scores spanning the entire scale used for the test. Thus the studies of TOEFL agree that the test is useful in discriminating English-speaking ability among non-native speakers. Clearly, the nature and level of TOEFL preclude direct translations of the scores into scores on verbal aptitude measures commonly used for selection of native-speaking students. The question remains, however: What is the relationship between tests which tap these differing aspects of verbal performance?

The present study examines the performance of non-native speakers of English on TOEFL and on some verbal aptitude tests designed for native speakers. For graduate-level students, the aptitude measure used was the verbal portion of the Graduate Record Examinations Aptitude Test (GRE-V). For undergraduates, two tests were included: the verbal portion of the College Board Scholastic Aptitude Test (SAT-V) and the Test of Standard Written English (TSWE). The two major ways of describing relationships between TOEFL and the verbal aptitude instruments are, first, to examine relative levels of performance of native and non-native speakers on the same test, and, second, to investigate the nature of the relationships between performance, on TOEFL and on the verbal aptitude test, by non-native

speakers. Both approaches are reported in this study. How the two groups compare is of interest, but differences in level are to be expected of the two populations. The second approach, however, has important practical implications. As part of the review process for admission to United States colleges and universities, scores on TOEFL and on GRE-V or SAT-V are frequently evaluated for foreign applicants who are not native speakers of English. For students whose proficiency in English (as measured by TOEFL) approaches that of native speakers, there would seem to be little problem in interpreting verbal aptitude scores. But for students who score below this level, English-language proficiency may play a significant role in their ability to cope with verbal aptitude tests written in English. Information that assesses the effect of the English-language factor in verbal aptitude tests and that provides some guidance on how results on one type of test can help interpret results on the other would thus be of considerable interest to those who must decide on the admission of foreign students to U.S. institutions of higher learning.

This report first describes the procedures used in the study and then presents the basic findings on the candidates' test performance, divided into undergraduate and graduate categories. With the performance data, we present the analysis of test results, including comparative information about native vs. non-native performance on the verbal aptitude tests. Included in this analysis are the means, standard deviations, and reliabilities, as well as the intertest and intratest correlations. The last portion of the analysis section documents th rformance of non-natives on TOEFL in relation to their performance on the verbal aptitude tests.

The appendix to this report summarizes a review, by a panel of specialists in English as a Second Language, of items from all the tests used in the study. The purpose of this review was to elicit expert judgments on the differences among the tests. As can be seen from the results, the review also stimulated judgments on the relative difficulty of the separate tests and suggestions for improving the appropriateness of TOEFL as a measure of English-language proficiency.

PROCEDURES

Test Selection

The first step was to select appropriate measures. The SAT and GRE verbal aptitude tests and the Test of Standard Written English seemed obvious choices. Nevertheless, it was felt that prior confirmation of what tests are currently used by academic institutions to screen foreign applicants would be advisable. Thus a telephone survey of admissions officers at 50 U.S. colleges and universities was conducted. The institutions were selected to provide a representative sample on the basis of size, geographical distribution, and category (public, private, etc.). Of the 50 institutions surveyed, four offer only bachelor's degrees. All the institutions are accredited and have a student population larger than one thousand. Table 1 summarizes the survey data. For admission of undergraduate students, 42 of the schools require foreign applicants to take TOEFL. Eight of the 42 that require TOEFL accept the substitution of the Michigan English Test or a course in English as a Second Language (ESL). Twelve of the 50 institutions require the SAT, but four accept the substitution of the American College Testing program (ACT). Only two require foreign applicants to take College Board Achievement Tests. All

those who take SAT also take TSWE, since it is administered along with the SAT.

See Table 1 on page 27

Of the 46 institutions that offer graduate-level programs, six noted that admissions are handled exclusively by the various academic departments; thus no single policy applies to all. Of the remaining 40 schools offering graduate degrees, thirty-six require TOEFL and only three allow a substitute. Twenty-four require foreign applicants to take the GRE-V. Only one school requires GRE but not TOEFL. Thirteen graduate schools require some applicants to take the Graduate Management Admission Test (GMAT) instead of GRE. Four schools require the Miller Analogies Test, and one requires the National Teacher Examinations (NTE) for foreign applicants. Table 2 summarizes the principal data from the graduate survey. The results of the survey led us to conclude that SAT-V, GRE-V, and TSWE were indeed appropriate instruments for the experimental portions of the study and that, although the relationship of language proficiency and verbal aptitude is of particular importance to graduate schools, sufficient numbers of undergraduate institutions used both measures to justify our also including an undergraduate sample.

See Table 2 on page 27

Design Considerations

The analysis of student performance on the designated tests was not based on data from already existing files. We believed that our purposes could best be met by arranging new test administrations, since we would avoid the problem of large time lapses between administrations of the two

tests and thus possible changes in English proficiency arising during that interval. Although this approach increases the precision with which the relationships may be described, it should be kept in mind that motivational factors might differ from those in effect in operational administrations of the tests. Admissions decisions were not being made on the basis of these aptitude scores, and thus perceived pressure to perform well may have been lower. Furthermore, some degree of self-selection operated among those agreeing to take the aptitude tests. Therefore, the relationships reported here should be taken as provisional, pending conversion findings based on operational administration of the tests.

Sample Selection

After test centers were identified where supervisors agreed to give the experimental tests (GRE-V, SAT-V, and TSWE) in the afternoon following the regular morning TOEFL administration, candidates were asked if they would participate in the study. Approximately 600 candidates were approached, equally divided between those applying to undergraduate and graduate institutions. Of these, 415 students agreed to participate, took TOEFL in the morning, and returned to take one of the experimental tests in the afternoon. Because of an irregularity in test administration at one center, some of the scores could not be used in the experiment. As a result, a group of 210 undergraduate-level students and a group of 186 graduate-level students were available for study.

The following data, based on responses to questions asked on the day of the test, describe the experimental groups:

Sex. Each group, undergraduate and graduate, is about 65 percent male, 35 percent female.

Age. The median age is 29 for the graduate group and 21 for the undergraduate group. The spread of ages is greater for the graduate-level group. The median age category of the graduate group extends from 25 to 32 years, whereas the median age category of the undergraduate group extends from 19 to 24 years.

Years of English study. The median number of years of English study is 7 for the graduate group, 6 for the undergraduate group.

Months in the U.S. In response to the question "How long have you been in the United States?" the graduate-level group reported an average of 13 months and the undergraduate group an average of 9 months.

Language spoken outside of class. The participants were asked to indicate whether they usually spoke English or their native language outside of class. About 60 percent of each group marked "Native Language," and about 32 percent marked "English." The remainder marked both or did not respond to the question.

Native country. Forty-two different native countries were listed by the graduate group, and fifty were listed by the undergraduate group. The largest number from the same country is 30 (i.e., 16%) for the graduates and 45 (i.e., 21%) for the undergraduates. The eight largest national groups in each sample are listed in Table 3.

See Table 3 on page 28

Native language. Thirty-five different languages were listed by the graduate group and thirty by the undergraduates. The largest group speaking a single language was 26 (Farsi, 14%) for the graduates and

47 (Farsi, 22%) for the undergraduates. The ten largest language groups in each sample are listed in Table 4.

See Table 4 on page 28

We did not wish to reveal the identity of the GRE, SAT, and TSWE; therefore, in all test booklets used and in all correspondence with students and supervisors, the tests were referred to as "Experimental Test-Graduate Level" and "Experimental Test-Undergraduate Level."

Test Administration

All 396 students who agreed to participate in this study, at thirteen test centers throughout the United States, took two tests. All took TOEFL in the morning, following the normal procedures for that testing program. In the afternoon, the 186 graduate applicants returned to take the graduate-level experimental test (GRE-V), and the 210 undergraduate applicants took the undergraduate-level experimental test (SAT-V and TSWE).

RESULTS AND DISCUSSION

In analyzing the results of the tests taken by the subjects in this study, we initially considered the representativeness of the two groups, graduate and undergraduate, in relation to the TOEFL population as a whole. This could best be checked by comparing the performance on TOEFL of the two groups of non-native speakers who participated in this study with the performance of a representative group of other non-native speakers who took the same form of the test on the same date in May 1977. Secondly, to address the major questions raised in this study, we compared the performance of the two groups on the respective "other" test(s) with the performance of

native speakers who took the same forms of the tests. To this end, the graduate and undergraduate groups were analyzed separately.

In order for us to compare the performance of the non-native subjects across tests, it was useful to look at basic statistical data for the three pairs of tests: TOEFL and GRE-Verbal, TOEFL and SAT-Verbal, and TOEFL and TSWE. In this portion of the discussion we therefore include means and standard deviations as well as correlation coefficients between TOEFL and the other tests. The overall distributions across tests are presented as scatterplots for the two groups involved; they provide information on how scores on one test can be used to interpret scores on the other.

Representativeness of the Sample

If the performance of the experimental groups were to indicate that they did not represent the typical population of non-native English speakers who take TOEFL, any analysis and interpretation of results from this study would be of questionable generalizability. In fact, the score distributions of both the graduate and undergraduate groups that participated in the study were reasonably representative of the general TOEFL population, although they were somewhat higher. This conclusion can be derived from a direct comparison of the two groups with other non-native speakers who took the same form of TOEFL on the same date in May 1977. Of the total number of 6,291 such persons participating in that administration at centers around the world, a representative sample of 1,540 cases was used to compile test data for that form of TOEFL. As shown in Table 5, both the graduate and undergraduate experimental groups performed better than the statistical sample did. The mean score for the experimental graduate group was a full 30 points higher than mean scores for the sample (523 vs. 493).

The undergraduate mean was 9 points higher than the sample's (502 vs. 493). Although to a lesser extent, the mean scores achieved by both experimental groups for each of the three sections reported by TOEFL were also higher than the corresponding mean scores for the sample.

See Table 5 on page 29

Further evidence of how the groups compare with other TOEFL candidates can be found in the data from all administrations of the test, worldwide, conducted from September 1976 to May 1977. For 50,072 graduate students in this category, the mean score was 506. For 44,149 undergraduates, the mean score was 502. Again we find that our two experimental groups can reasonably be considered as representative of TOEFL candidates, the undergraduate group being the more representative of the two, as would be consistent with their shorter average period of residence in the U.S.

Native vs. Non-native Comparisons

The next question concerns the performance of the experimental groups on tests other than TOEFL. Looking first at the data for the graduate group, we would do well to recall that GRE-V was not designed, as TOEFL was, to measure English proficiency, nor was it designed with non-native speakers in mind. Thus, it should not be expected that effective comparisons of proficiency could readily be made of groups or of individuals who had taken GRE-V and TOEFL. Nevertheless, since many non-native English speakers do take the GRE Aptitude Test and subsequently must have their verbal scores reviewed by admissions offices and academic departments, it

is helpful to see how their performance compares with that of native speakers who take the same tests.

In this case, there is no control group representing native speakers who took the GRE test and TOEFL at the same time. The data used for comparison here are taken from an analysis of the performance of a representative sample of 1,495 native speakers who took the same form of the GRE Verbal Aptitude test at the same time (May 1977).

As can be seen in Table 6, the graduate candidates in the non-native group, although they were typical of the TOEFL population, scored much lower on GRE-Verbal than the native-speaking sample did.

See Table 6 on page 30

Scores from GRE-Verbal are reported on a 200-900 scale. Within this range, the native speakers had a mean score of 514. The non-natives, however, had a mean score of only 274. Clearly, scores that cluster near the bottom of the scale do not lend themselves to easy interpretation, particularly in a multiple-choice testing situation, in which blind guessing yields an expected score of 200, with a standard deviation of about 70. The primary conclusion we can draw from these results is that GRE-V is far too difficult for most non-native speakers of English.

No figures are given here for subscores since only total scores are reported for GRE-V. This test does, however, contain two different types of item. Of the 100 items in GRE-V*, 60 measure verbal reasoning,

*In October 1977, the GRE Aptitude Test was restructured. The Verbal section was reduced from 100 items to 80 items, and it now is timed for 50 minutes rather than 75. However, scores on the new and old format are comparable.

including analogies, antonyms, and sentence completions. The remaining 40 are reading comprehension items requiring the candidate to respond to a variety of questions based on prose passages. To obtain precise information on how the non-native subjects compared to the native sample in their ability to cope with the two separate categories of items, we looked at raw scores rather than scaled scores. Here we are referring to the actual number of items correctly answered on the test. Table 7 indicates the means and standard deviations for the native and non-native groups on each of the two subparts of GRE-Verbal. The difference between the two groups is, for all practical purposes, the same for the verbal reasoning and the reading sections. The data do confirm the earlier evidence, in the form of subpart means, that native speakers performed much better than non-native.

See Table 7 on page 30

One further consideration in comparing the performance of the two groups on GRE-V is that of speededness. Once again it was useful to look at the two sections for the test separately. In fact, we found a greater difference between the groups when we separated the sections. In Table 8 we note that the speededness factor appears to have a similar effect on both groups in the set of 60 verbal-reasoning items. A much clearer difference appears, however, in the set of 40 reading items. Even for native speakers of English, GRE-V is speeded in the sense that a fairly large number of candidates do not complete the test. But in these reading comprehension items, non-native speakers seemed to have even greater difficulty in completing the test than did the native speakers.

See Table 8 on page 30

A possible factor in the performance difference discussed above is the effect of the reading load on non-native speakers. Although the non-native speakers do not seem to require any more time than the native speakers do to process discrete items like those contained in the verbal analogies or antonyms, this does not appear to be the case for the reading comprehension items. No definite conclusions can be drawn on this point as a result of this study. However, the data shown here do point to a variable that could well be significant in all considerations of non-native speakers' performance on tests oriented to native speakers.

At the undergraduate level, similar comparisons were made. Here our point of comparison was the group of 232,021 native speakers who took the same form of the Scholastic Aptitude Test in December 1974. Data from a representative sample of 1,765 candidates were used to make the native-non-native speaker comparison. Since the SAT-V and TSWE were administered during the same administration to both the non-native group that participated in the study and the native-speaker group, the results of both tests are included. Table 9 displays the summary test data for the native and non-native groups.

See Table 9 on page 31

The verbal section of the SAT is reported on a 200-800 scale, the GRE-V is on a 200-900 scale, and the results indicate a relationship between the native and non-native undergraduates similar to that of the corresponding graduate groups. The mean score of 269 achieved by the

non-native group on SAT-V reveals once again that this group found the test very difficult, since the scores cluster near the low end of the scale. The principal difference between the undergraduate and graduate groups is in the mean score for the native speakers. Since the undergraduate native speakers achieved a mean score of 425 on SAT-V as opposed to the 514 mean score achieved by the graduate native speakers on GRE-V, the non-native undergraduates in our study appear to be not so far below (1.5 standard deviations) their native-speaking counterparts as the graduate group was (almost 2 standard deviations).

Two observations must again be stressed about these verbal aptitude tests. They are measures of ability to do undergraduate or graduate work, not a language proficiency test in the sense that TOEFL is. Again, neither verbal aptitude test is designed for non-native speakers. Both considerations must thus be kept in mind when interpreting these results.

TSWE, whose scale ranges from 20 to 60+, is used for placing entering college students in appropriate freshman English classes. It is a language test and more closely approximates TOEFL than does either GRE-V or SAT-V. Nevertheless, a large discrepancy remains between the native and non-native speakers with regard to their performance on TSWE. Quite probably, the results should not be interpreted in the same way for both groups. With reference to Table 9, it is important to note that the reliability of TSWE is very nearly the same for both groups. This was not true for either the SAT-V or GRE-V. That both of the latter tests exhibited low reliability for the non-native groups is important to consider when we make our overall comparisons of test performance.

When we compiled data on the tests, subscores were not identified for the SAT group. Therefore, it is not possible to discuss any differences that may exist between performance on discrete vocabulary items vs. reading. For the question of test speededness, however, data are available to compare the non-native group's speed in coping with the SAT-V and TSWE with that of the native-speaking group. Table 10 shows the comparative figures for both groups. The two sections of the SAT represent separately timed sections, each containing both vocabulary and reading items. The first section contains 45 items. The second contains 40 items. What is significant about these data is that, by usual measures of speededness, the non-native speakers encountered little more difficulty in meeting the time requirements of the test than the native speakers did. It is clear that the test is speeded for both groups. For the TSWE there is slightly more of a difference between the groups, at least in the percentage of candidates completing the test (75% of the natives vs. 65% of the non-natives). Even though TSWE is more closely related to TOEFL (at least to some of its sections) with regard to test content, completing the 50 items in the time allotted apparently introduces more speed demands on non-native than on native speakers.

See Table 10 on page 31

Test Relationships: Non-natives

To this point, the discussion and analysis of test results have focused on how the performance of experimental groups of non-native speakers on the "other-than-TOEFL" tests compared with that of native speakers on the same forms of those tests. The principal conclusion was that these tests are

difficult for non-natives and that, because their scores tend to cluster in the low ranges of the scales used by those tests, interpretation of scores could be complicated.

We turn our attention now to the relationship between TOEFL and the other tests by looking at the correlations among them. Here we concern ourselves only with the two non-native groups of graduates and undergraduates who participated in the study. Table 11 gives the data related to TOEFL and GRE-V for the graduate group.

See Table 11 on page 32

The overall correlation coefficient of .645 between TOEFL and GRE-V would seem to indicate that the two tests are to some extent related but are by no means identical in the skill being tested. If the part scores are considered, one additional point appears noteworthy. The listening comprehension section of TOEFL shows the lowest correlation with the GRE, a finding to be expected, since listening comprehension skills are not tapped in either of the two parts of GRE-V. The point worth noting is that in TOEFL the listening comprehension section shows a similar relationship to the other sections. No major difference appears in the relationship of the other two sections of TOEFL to GRE-V.

Looking at the correlation coefficients between TOEFL and the undergraduate tests, we find evidence of an increasing relationship. The .681 correlation coefficient between TOEFL and SAT-V totals (see Table 12) is slightly higher than that found between TOEFL and GRE-V. Similarly, the .720 correlation coefficient between the TOEFL total and TSWE is indicative of a closer relationship between those two tests than between

TOEFL and either of the two aptitude tests. This follows from TSWE's being a test of language ability, particularly of written English. Support for this assertion can be found in the fact that the highest part-score correlation coefficient between TOEFL and TSWE is that for the second section of TOEFL and TSWE (.708). The items used in this section of TOEFL most closely resemble those used on TSWE. From a similar point of view, it is the third section of TOEFL which shows the greatest relationship to SAT Verbal. Here the reading and vocabulary items in that section of TOEFL resemble the format of items used in SAT-V. As with GRE-V and TOEFL, the listening comprehension section of TOEFL shows the lowest relationship to either TSWE or SAT-V.

See Table 12 on page 32

Overall Test Comparisons

The principal method of describing the performance of the experimental groups has been to present the correlations between the TOEFL scores and scores on the other graduate or undergraduate measures. A difficulty, however, is the radically different performance of the groups on the two tests. In this section we explain the nature of the statistical problem and describe certain statistical procedures we adopted to explore this problem and to go beyond these correlation coefficients. At the same time, we have felt the need to provide data that can support some broader comparative statements about how the tests in question are related. Thus, by means of the scatterplots presented here, we are able at least to make some tentative claims about how TOEFL scores may be used to identify thresholds of relationships to scores on GRE-V and SAT-V.

The undergraduate results can be used to exemplify the problem. If one examines the reliabilities and intercorrelations for this group, it is apparent that, although TOEFL and TSWE are within a comparable range, the difficulty of SAT-V was inappropriate for measurement in most of the non-native speaking group. Although the reliablity for the TOEFL test was an encouraging .94, the total SAT reliability for this group was only .77. With raw-score means in the lower one-third of U.S. descriptive statistics, the experimental group is near the lower extreme of the SAT scale. The standard deviation for the non-native speaking group is less than two-thirds that for the English-speaking SAT sample. The Kuder-Richardson 20 reliability estimate is related to the range of scores in a group, and it is lower in more homogeneous samples. Although standard errors of measurement are similar across groups, this restriction of variation depresses correlations. The correlations obtained may therefore reflect these differences in relative difficulty and range of the instruments for the groups, the floor effects on SAT-V attenuating possible relationship to TOEFL. In spite of the distinct difficulties and the restricted variation of the experimental sample, the correlation of .68 between TOEFL total and SAT-V is substantial.

Figure 1 gives a scatterplot of TOEFL and SAT-V scores, revealing the characteristic pattern of two tests of widely dissimilar difficulties. If the underlying relationship between true scores is linear, as in Figure 2, but one or both tests are truncated at one end of their range, the resulting observed relationship appears triangular or curvilinear, as in Figure 3.

Although information is irretrievably lost when really different abilities at the lower levels of the SAT range are mapped into a small range of "chance level" scores, transformations of the scales can partially

straighten out the artificial curvilinearity or piecewise linearity induced by the distinct difficulty levels of the tests. While they do not retrieve the value of the correlation that would have been obtained in the absence of the floor effect, if such transformations increase the correlation, it may be concluded that the true relation is higher than the obtained one.

In the undergraduate sample, TSWE was of appropriate difficulty, yielding a nearly linear relationship (Figure 4); transformation of its scale would not be expected to increase the relationship. However, transforming scales for the TOEFL-SAT-V relationship might be expected to increase correlations. Two transformations were investigated: log SAT-V vs. TOEFL, and a correlation based on only those cases corresponding to median SAT-V scores above the chance level. The log transformation has the effect of "squeezing" the upper portion of the SAT scale and of tending to straighten out nonlinear relationships which exhibit increasing slope. The latter approach, "truncation," or trimming, corresponds to a piecewise linear fit, discarding those cases in the OP region of Figure 3, and fitting only those cases clustering around line PQ.

As expected, neither the log nor the truncation transformations increased the TSWE-TOEFL correlations. Table 13 shows the correlations of observed scores, log TSWE vs. TOEFL and TSWE vs. truncated TOEFL.

See Table 13 on page 36

The log transformation has essentially no effect on correlations with TSWE, and truncation actually decreases the correlations. In contrast, as shown in Table 14, the transformation does yield very small increases in

correlations between TOEFL I, II, and Total and the SAT-V scores, as does truncation for TOEFL I and TOEFL Total. Neither transformation increases the correlation of TOEFL II--structure and written expression--with SAT-V.

See Table 14 on page 36

These results suggest that the true relationship between SAT-V and both the TOEFL listening and the TOEFL reading and vocabulary subtests is somewhat higher than the observed relationship would indicate but that the relationship of the TOEFL structure and written expression subtest to SAT-V is not underestimated by observed score correlations. However, the changes in correlations are too small to be of practical significance.

In the graduate population, neither the discrete verbal nor the reading comprehension sections of GRE-V yielded raw score means as high as 12% of the total number of items for the non-native English-speaking samples. These scores are well below those to be expected by chance if candidates had attempted all items. In this situation, we would expect scale transformations to lead to increased correlations between TOEFL and the total GRE verbal scores.

Table 15 shows that these expectations are confirmed. In this table, two additional transformations ($\frac{1}{\text{GRE-V}}$ vs. TOEFL and GRE-V vs. TOEFL^3) are introduced in an attempt to straighten the marked curvilinearity apparent in Figure 5. These transformations have effects similar to the log transformation but by the route of stretching the TOEFL scale (extreme stretching for TOEFL^3--which raises TOEFL scores to the third power) and are applicable to curvatures even more pronounced than those which may be rectified by the logarithmic transformation.

See Table 15 on page 38

Here we see a regular increase with the rather extreme stretching of the top portion of the TOEFL scale obtained by cubing TOEFL scores yielding the greatest increase in correlations.

Examining the observed score and TOEFL3 correlations with the two GRE verbal subsections, vocabulary and reading, as shown in Table 16, we see a similar pattern. There is a greater improvement through scale transformation among correlations of the discrete verbal reasoning items in GRE I with TOEFL. These are most pronounced for either GRE subscore with TOEFL III--reading comprehension and vocabulary.

See Table 16 on page 38

It is possible that more extreme transformations would further increase the correlations, but the point that floor effects on the GRE distort the true relationship has been amply documented.

For TOEFL reading and vocabulary and the total, truncation does as well as cubing the TOEFL score, and examination of the points of truncation offers a rough estimate of the minimum TOEFL score at which GRE scores begin to become interpretable. Table 17 represents the estimated truncation points (corresponding to point P in Figure 3), or those TOEFL scores at which GRE scores begin to rise from their floor and to exhibit positive correlations with TOEFL scores.

These truncation points might be thought of as minimal values of TOEFL scores for which it makes sense to examine a candidate's GRE verbal score.

See Table 17 on page 39

The information in this table may be summarized by suggesting that below a TOEFL score of about 475, differences in GRE verbal scores are unlikely to be interpretable. Following similar procedures, we can suggest that SAT verbal scores are unlikely to be informative below TOEFL scores of about 435. Here the TOEFL cut score is lower, suggesting that SAT verbal scores are likely to be interpretable for a larger proportion (perhaps 80%) of the TOEFL undergraduate candidates. Interpretability does not mean equivalence, however, and even the TOEFL-TSWE correlations, least distorted by floor effects, show that the two tests share only 52% of their variance, thus suggesting that the instruments are far from interchangeable.

GENERAL CONCLUSIONS

A number of important conclusions can be made as a result of this study. It is clear, first of all, that non-native English speakers do not perform as well on the GRE and SAT verbal aptitude tests or on the TSWE as they do on TOEFL. This was to be expected given the nature and purpose of these tests. The data provided here, however, do show how non-native speakers compare with native speakers in performance on three tests other than TOEFL. With this information, interpretations of score reports from these tests can be more easily made.

In this regard, the most useful result of the study was the identification of score levels on TOEFL at which scores on the other tests begin to be meaningful. This information would, of course, be useful only for students who have taken both TOEFL and the other test(s). But since most

foreign students who apply to colleges and universities fall in this category, the information should be valuable for the admissions process.

The review of test items by the panel of specialists in English as a Second Language does not of course, constitute experimental data (see Appendix). Nevertheless, the comments by the reviewers provided important supporting information for the future refinement of TOEFL. In particular, the comments on the length and nature of the reading passages and related items used in TOEFL and the other tests have yielded valuable information for future test construction.

The most significant result of this study, which relates to both the item review and candidate performance, is the manner in which both coincide. Regarding overall performance, the GRE verbal test proved to be the most difficult for the non-native speaker candidates. The next most difficult was the SAT verbal, and third was TSWE. Looking at the comments of the reviewers, we find that their order of preference for items is exactly the same.

This study represents the first significant attempt to compare performance on TOEFL with that on tests like the verbal aptitude tests included here. All the conclusions reached in this study should prove useful for interpreting foreign student performance. Additional studies will no doubt raise more specific questions and attempt to reach even more practical conclusions than was possible in this study.

REFERENCES

Angelis, P. J. "Language Testing and Intelligence Testing: Friends or Foes?" Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing. Carbondale, Ill.: Southern Illinois University, 1977. Pp. 97-104.

Angoff, W. A., & Sharon, A. T. A Comparison of Scores Earned on the Test of English as a Foreign Language by Native American College Students and Foreign Applicants to United States Colleges. ETS Research Bulletin 70-8. Princeton, N.J.: Educational Testing Service, 1970.

Clark, J. L. D. The Performance of Native Speakers of English on the Test of English as a Foreign Language. TOEFL Research Report Number 1. Princeton, N.J.: Educational Testing Service, 1977.

TABLES AND FIGURES

TABLE 1

UNDERGRADUATE-LEVEL ADMISSIONS TESTS

REQUIRED FOR FOREIGN APPLICANTS

Test	Number of Institutions	Percent
TOEFL	42[a]	84%
SAT	12[b]	24%
CB-Achievement	2	4%

[a]Eight institutions accept the Michigan English Test or an ESL course in place of TOEFL.

[b]Four institutions accept ACT in place of SAT.

TABLE 2

GRADUATE-LEVEL ADMISSIONS TESTS

REQUIRED FOR FOREIGN STUDENTS

Test	Number of Institutions	Percent
TOEFL	36[a]	90%
GRE-V	24[b]	60%

[a]Three institutions accept a substitute test or ESL Course for TOEFL.

[b]Thirteen institutions require GMAT instead of GRE for some applicants.

TABLE 3

NATIVE COUNTRIES AND NUMBERS OF PARTICIPANTS
(Eight largest groups)

Graduate Group		Undergraduate Group	
Country	N	Country	N
India	30	Iran	45
Iran	24	Hong Kong	24
Philippines	17	Japan	16
Korea	11	Vietnam	11
Vietnam	10	Indonesia	9
Japan	10	China	8
China	10	Nigeria	8
Thailand	9	Korea	7
Other countries	65	Other countries	82
Total	186	Total	210

TABLE 4

NATIVE LANGUAGES AND NUMBERS OF PARTICIPANTS
(Ten largest groups)

Graduate Group		Undergraduate Group	
Language	N	Language	N
Farsi (Persian)	26	Farsi (Persian)	47
Gujarati	14	Chinese	36
Spanish	13	Spanish	17
Chinese	12	Japanese	16
Arabic	12	Arabic	16
Korean	11	Vietnamese	11
Japanese	10	Indonesian	9
Vietnamese	10	Korean	7
Tagalog	10	Greek	7
Thai	9	Yoruba	5
Other languages	59	Other languages	39
Total	186		210

TABLE 5

TOEFL COMPARATIVE DATA

	EXPERIMENTAL GROUP GRADUATE (n=186)				COMPARISON GROUP (n=1,540)				EXPERIMENTAL GROUP UNDERGRADUATE (n=210)			
	MEAN	S.D.	RELIAB.	S.E.M.*	MEAN	S.D.	RELIAB.	S.E.M.*	MEAN	S.D.	RELIAB.	S.E.M.*
I. LISTENING COMPREHENSION	53.72	7.03	.89	2.9	52.07	7.43	.89	2.9	53.85	6.52	.86	2.4
II. STRUCTURE AND WRITTEN EXPRESSION	50.60	7.99	.83	2.7	47.32	8.71	.86	2.7	48.11	7.63	.81	2.4
III. READING AND VOCABULARY	52.70	7.85	.91	3.1	48.53	8.56	.92	3.3	48.66	7.40	.89	3.3
TOTAL	523	69	.95	15	493	75	.95	17	502	63	.94	16

*Standard Error of Measurement

TABLE 6

GRE-VERBAL SCORE COMPARISONS

	MEAN	S.D.	RELIAB.	S.E.M.*
EXPERIMENTAL GROUP -- NON-NATIVES (n=186)	274	67	.78	30
NATIVE SPEAKERS (n=1,495)	514	128	.94	32

TABLE 7

GRE SUBPARTS

	VERBAL REASONING				READING			
	MEAN	S.D.	RELIAB.	S.E.M.*	MEAN	S.D.	RELIAB.	S.E.M.*
EXPERIMENTAL GROUP -- NON-NATIVES (n=186)	5.10	7.15	.69	3.5	3.87	4.60	.47	3.0
NATIVE SPEAKERS (n=1,495)	27.61	11.99	.92	3.5	18.49	8.55	.84	3.4

TABLE 8

TEST SPEEDEDNESS

	VERBAL REASONING		READING COMP.	
	Natives	Non-natives	Natives	Non-natives
Per cent completing test	84.5	48.9	61.3	47.2
Per cent completing 75% of test	91.4	84.4	94.7	75.6
Number of items reached by 80% of candidates	53	49	35	27
Total # of items	60 items		40 items	

*Standard Error of Measurement

TABLE 9

SAT AND TSWE SCORE COMPARISONS

	SAT VERBAL				TSWE			
	MEAN	S.D.	RELIAB.	S.E.M.	MEAN	S.D.	RELIAB.	S.E.M.
EXPERIMENTAL GROUP NON-NATIVE (n=210)	269	67	.77	33	28	8.8	.84	4
NATIVE (n=1,765)	425	106	.91	32	42.35	11.09	.89	3.7

TABLE 10

TEST SPEEDEDNESS

	SAT I		SAT II		TSWE	
	Native	Non-native	Native	Non-native	Native	Non-native
Per cent completing test	72.5	73.5	74.5	65.5	75.4	65.0
Per cent completing 75% of test	99.2	98.5	97.4	90.5	96.0	89.5
Number of items reached by 80% of candidates	42	41	39	38	47	41
Total # of items	45 items		40 items		50 items	

TABLE 11

TOEFL-GRE INTERCORRELATIONS

TOEFL (n=186)

	LISTENING COMPREHENSION	GRAMMAR AND WRITTEN EXPRESSION	READING COMPREHENSION AND VOCABULARY	
SECTION*	I	II	III	TOTAL
I. LIST. COMP.	----	.698	.723	.878
II. STR & WE	.698	----	.801	.922
III. RC & VOC	.723	.801	----	.924
TOEFL TOTAL	.878	.922	.924	----
GRE-V TOTAL	.521	.612	.623	.645

TABLE 12

TOEFL-SAT-TSWE INTERCORRELATIONS

TOEFL (n=210)

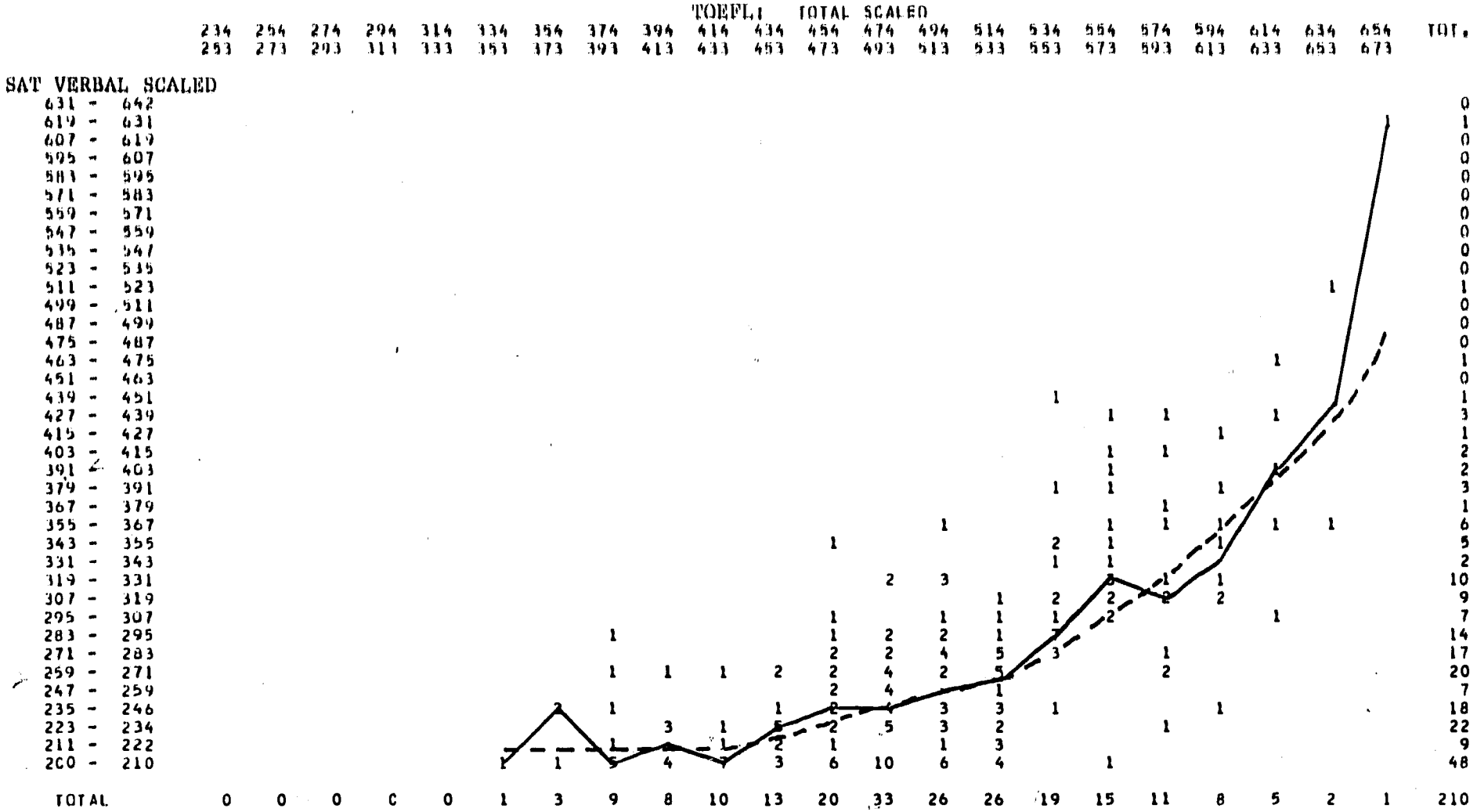
SECTION*	I	II	III	TOTAL
I. LIST. COMP.	----	.537	.633	.810
II. STR & WE	.537	----	.769	.890
III. RC & VOC	.633	.769	----	.920
TOEFL TOTAL	.810	.890	.920	----
SAT-VERBAL	.449	.643	.681	.681
TSWE	.512	.708	.657	.720

*Section I: Listening Comprehension; Section II: Structure and Written Expression; Section III: Reading Comprehension and Vocabulary.

FIGURE 1

TEST OF ENGLISH AS A FOREIGN LANGUAGE

UNDERGRADUATE-LEVEL SAMPLE



GROUP	N	MIN	MAX	MEAN	SD N	SD N-1
TOEFL: TOTAL SCALED	210	350.0000	663.0000	502.1379	62.7676	62.9176
SAT VERBAL: III SCALED	210	200.0000	620.0000	269.2476	66.6509	66.8102

0.6810

——— Medians
- - - - Smoothed Medians

FIGURE 2

Underlying Relationship Between
Abilities Measured by Tests A & B

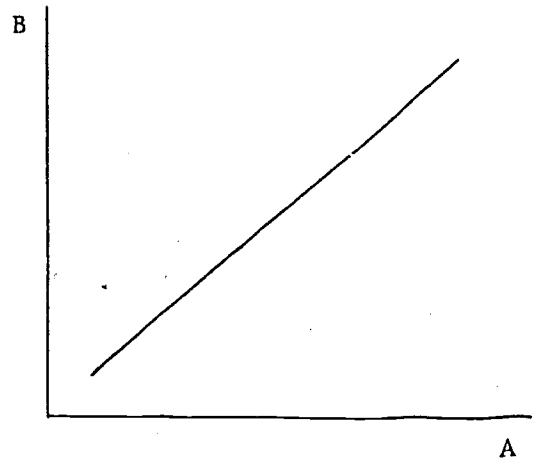


FIGURE 3

Observed Relationship When B' is an
Overly Difficult Measure of the Trait
Measured by B

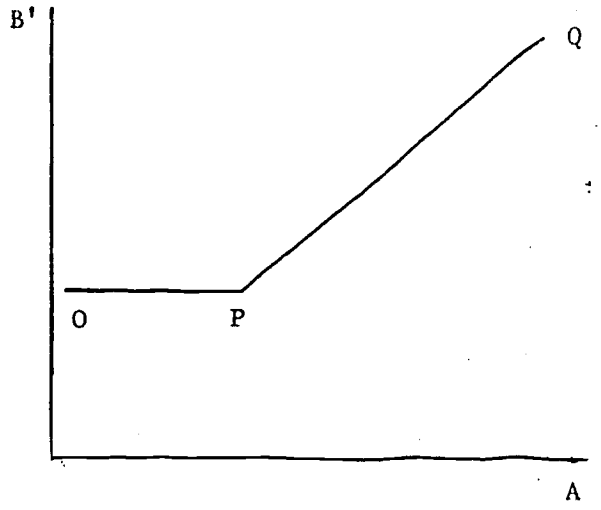
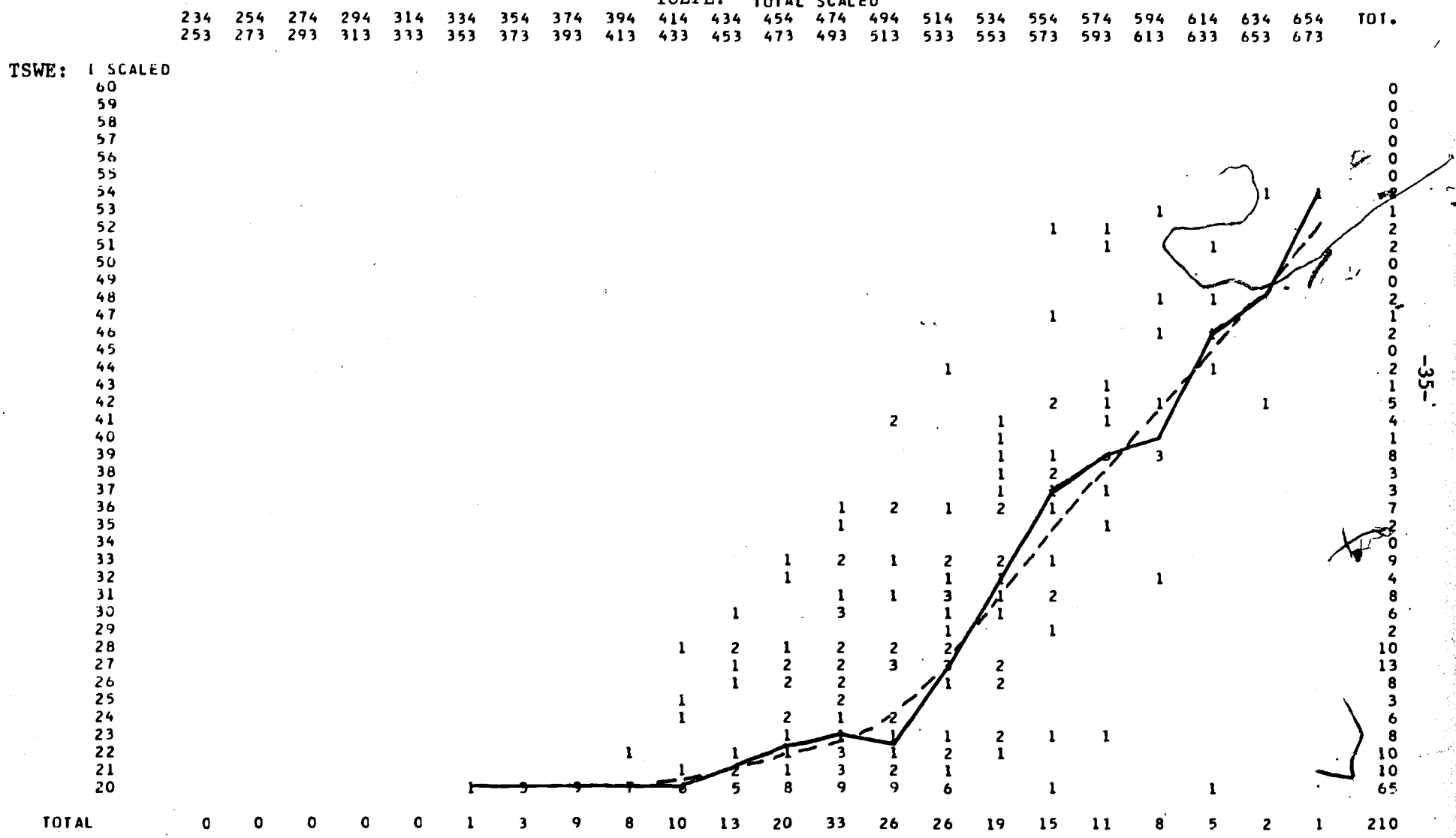


FIGURE 4

TEST OF ENGLISH AS A FOREIGN LANGUAGE

UNDERGRADUATE-LEVEL SAMPLE



GROUP		N	MIN	MAX	MEAN	SD N	SD N-1
TOEFL:	TOTAL SCALED	210	350.0000	663.0000	502.1379	62.7676	62.9176
TSWE:	I SCALED	210	20.0000	54.0000	28.0428	8.8249	8.8460
R=	0.7202						

——— Medians
- - - - Smoothed Medians

TABLE 13

OBSERVED AND TRANSFORMED CORRELATIONS

TOEFL AND TSWE

	TOEFL SECTION*			
TSWE	I	II	III	TOTAL
Observed	.512	.708	.657	.720
Log	.512	.703	.660	.718
Truncated	.434	.630	.601	.707

TABLE 14

OBSERVED AND TRANSFORMED CORRELATIONS

TOEFL AND SAT

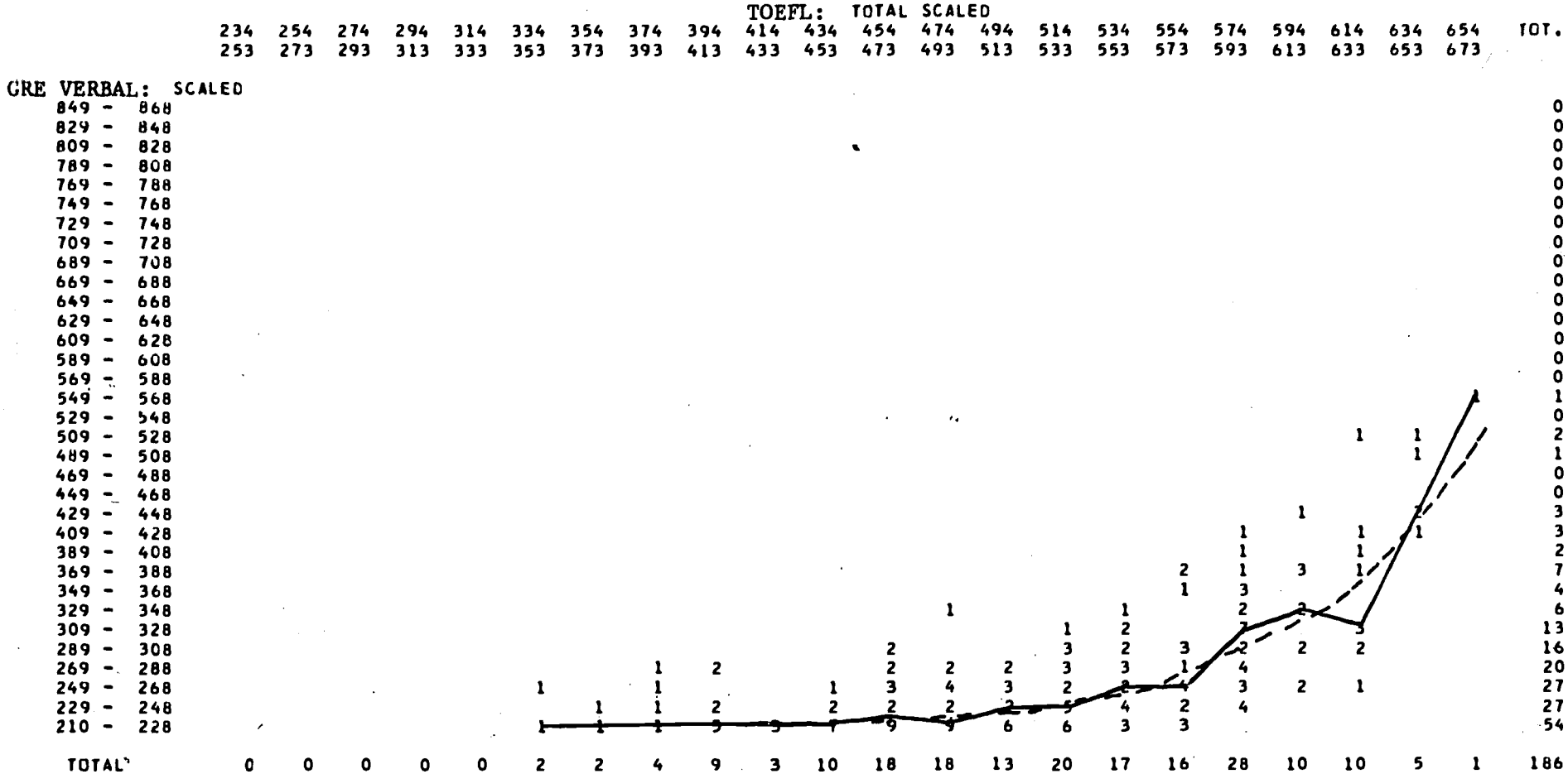
	TOEFL SECTION*			
SAT-V	I	II	III	TOTAL
Observed	.449	.643	.681	.681
Log	.450	.637	.690	.687
Truncated	.452	.636	.679	.687

*Section I: Listening Comprehension; Section II: Structure and Written Expression; Section III: Reading Comprehension and Vocabulary.

FIGURE 5

TEST OF ENGLISH AS A FOREIGN LANGUAGE

GRADUATE-LEVEL SAMPLE



GROUP	N	MIN	MAX	MEAN	SD N	SD N-1
TOEFL: TOTAL SCALED	186	337.0000	663.0000	523.3977	69.1972	69.3840
GRE VERBAL: I+II SCALED	186	210.0000	560.0000	273.5806	66.3221	66.5011

R= 0.6450

——— Medians

- - - - Smoothed Medians

TABLE 15

OBSERVED AND TRANSFORMED CORRELATIONS

TOEFL AND GRE

GRE-V	TOEFL SECTION I	II	III	TOTAL
Observed	.521	.612	.623	.645
Log (GRE-V)	.527	.618	.648	.662
$\frac{1}{1-\text{GRE-V}}$	.530	.616	.662	.663
Truncated	.533	.627	.684	.703
(TOEFL^3)	.560	.657	.686	.703

TABLE 16

GRE-VERBAL PART CORRELATIONS WITH TOEFL AND (TOEFL^3)

GRE I (verbal reasoning)	I	II	III	TOTAL
Observed	.487	.582	.604	.616
(TOEFL^3)	.529	.633	.666	.674
GRE II (reading)				
Observed	.453	.503	.499	.534
(TOEFL^3)	.482	.530	.548	.577

TABLE 17

TRUNCATION POINTS FOR TOEFL AND
GRE-VERBAL SCORES

TOEFL SCORE	TRUNCATION POINT
Section I	50
Section II	44
Section III	51
Total	474

APPENDIX

A-1

Appendix

Item Review

As indicated in the earlier section describing the procedures followed in this study, an attempt was made to gather information from a representative group of specialists in English as a Second Language on the relationship of the various tests administered. Ten specialists representing different ESL programs throughout the United States were chosen to review the tests. Because one of the ten was not able to complete the assignment, the data given here are the result of the reviews of nine ESL specialists. Despite the small number, it was felt that the group, chosen because of their longstanding familiarity with all aspects of ESL training, would represent ESL specialists in general.

The purpose of this review was to obtain the views of specialists on the similarities and differences among the item types found in the various tests. For this reason the specific tests were not identified. All the items from all four tests (TOEFL, GRE, SAT, and TSWE) were first divided by skill or area tested. Within each of the resulting four groups (Reading, Vocabulary, Writing, Listening), the items from the various tests were first divided into groups and then randomized. Along with the items to be reviewed, the specialists were asked to complete a questionnaire in which they were asked to indicate which items they felt were the most or least appropriate "for testing the English proficiency of non-native speakers who are being evaluated for admission to full-time academic (not ESL) study in American colleges and universities."

The responses of the reviewers can best be described for each separate section. For the first group (reading), none of the reviewers chose the

items taken from GRE-V as being appropriate for testing the reading proficiency of the groups described. Three chose the SAT reading items as the most appropriate, and the remaining six chose the TOEFL reading items. At first glance, these results would seem to be contrary to the actual performance of the non-native students who participated in the study. In the comments included by the reviewers, however, all six who chose the SAT items mentioned that those items seemed preferable to the TOEFL items because they contained longer, more realistic, reading passages for students who are to enter full-time academic study. Such comments are very pertinent in the light of the inclusion of the short, practical selections in the TOEFL reading section since the introduction of the three-part form of the test in 1976. In the last section of the questionnaire, in which the reviewers were asked to indicate their choice of the most appropriate items for testing reading, they showed some ambivalence about choosing between the TOEFL and SAT items. The reviewers felt that the level of difficulty of the TOEFL items was about what it should be but that the main weakness was their not regularly testing comprehension of extended passages.

In the second group of items (vocabulary), seven of the nine reviewers selected the TOEFL vocabulary items as the most appropriate. The principal argument given for this choice was the use by TOEFL of sentence-length contexts for testing the meanings of words. The verbal analogy type of item was felt to be too restrictive and only indirectly related to a subject's knowledge of the meanings of words.

The third group of items was entitled "Writing" and contained only two types of items from TOEFL (structure and written expression) and two types

from TSWE (not named as such but roughly similar to the TOEFL items). The choices indicated by the reviewers did not focus on a clear division between the TOEFL and TSWE items. The preference in the written expression item type (those items requiring recognition of an error in a given sentence) was clearly toward the TSWE items. But, again, a particular feature of the items explained the choice: that the TSWE uses five-choice items as opposed to the four-choice items used by TOEFL. The choices for the structure type of item were almost evenly divided between those from TOEFL and TSWE.

The last group included the items for testing listening comprehension. In this case, only TOEFL items were included because the other tests do not measure listening. The section was nevertheless included for review in order to cover all items in the tests used in the study and to get some feedback on the differences among the three types used in TOEFL. The choices were most in favor of the mini-talks (4), next for the dialogs (3), and least for the one-sentence rejoinders (2). On the whole, the comments on this section expressed a greater desire for those items that contained greater context. Also, the one-sentence items were considered to be less realistic than either of the other two for testing listening comprehension.

In summary, the choices of the group of specialists indicated a distinct preference for TOEFL items to test vocabulary, of TOEFL items and to some extent SAT items to test reading, and a combination of TOEFL and TSWE type items to test writing. Although this information was surely peripheral to the primary purpose of the study, the comments do provide valuable guidance on how TOEFL might best measure the English skills needed by foreign students entering U.S. colleges and universities.

This Research Report is part of a series of reports on research relating to the Test of English as a Foreign Language. Other reports include:

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

Discusses the results of the administration of TOEFL to native speakers of English just prior to their graduation from a college-preparatory high school program. Total test score distributions were highly negatively skewed, reinforcing findings of earlier studies that TOEFL is not psychometrically appropriate for discriminating among native speakers of English with respect to English language competence.

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

Describes an extensive research study conducted from 1972 to 1974 that was designed to explore possible changes in the format and content of TOEFL. Questions of validation, criterion selection, and content specifications were investigated. The report includes the results of these findings and discusses the implications for TOEFL content specifications and internal structure. This study contributed to the restructuring of TOEFL beginning in 1976.

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

Describes a three-year study involving the development and experimental administration of test formats and item types aimed at measuring the English-speaking proficiency of non-native speakers. Factor analysis and other techniques were used to identify subsets of item formats and individual items having satisfactory correlations with the Foreign Service Institute criterion interview administered to the test subjects. The results were grouped into a prototype "Test of Spoken English."

The above reports are currently available. Other research reports are planned. For further information about any of the TOEFL Research Reports, write to:

TOEFL Program Office
Box 899
Princeton, NJ 08541, USA