DOCUMENT RESUME

ED 065 537 TM 001 459

AUTHOR	Angoff, William H.; Sharon, Amiel T.
TITLE	Patterns of Test and Item Difficulty for Six Foreign Language Groups on the Test of English as a Foreign Language.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RDR-71-72-5
PUB DATE	Feb 72
NOTE	36p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Analysis of Variance; Data Collection; *English (Second Language); *Factor Analysis; Group Tests; *Item Analysis; Second Language Learning; Statistical Analysis; *Test Construction; Test Results; *Tests

ABSTRACT

Two-factor analyses of variance with multiple measurements on one factor were conducted for the five sections of the Test of English as a Foreign Language (TOEFL) and six language groups, and also for the items of each section, based on data collected at the first administration of this form of the test, in October 1969. the principal findings, that the Test x Group interaction and the Item x Group interactions were significant, indicated that some tests and some items were relatively more difficult for some language groups than for others. Displays of the Item x Group interactions were made by analyzing the item difficulty plots for each language group against a spaced sample of all candidates taking this form of the test in October 1969. A measure of the deviation of each item from the central tendency of the plot was developed, expressing the degree to which the item was especially difficult--or-easy--for a particular language group relative to the other items. Distributions of these measures are given for each of the five parts of TOEFL, for each of the six language groups. (Author)

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS
RDR-71-72, NO. 5

•

RESEARCH BULLETIN
RB-72-2 FEBRUARY 1972

Patterns of Test and Item Difficulty for Six Foreign Language Groups on the Test of English as a Foreign Language

•

William H. Angoff and Amiel T. Sharon

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

PATTERNS OF TEST AND ITEM DIFFICULTY FOR SIX FOREIGN LANGUAGE GROUPS ON THE TEST OF ENGLISH AS A FOREIGN LANGUAGE

William H. Angoff and Amiel T. Sharon

Educational Testing Service

Abstract

Two-factor analyses of variance with multiple measurements on one factor were conducted for the five sections of the Test of English as a Foreign Language (TOEFL) and six language groups, and also for the items of each section, based on data collected at the first administration of this form of the test, in October 1969. The principal findings, that the Test x Group interaction and the Item x Group interactions were significant, indicated that some tests and some items were relatively more difficult for some language groups than for others.

Displays of the Item x Group interactions were made by analyzing the item difficulty plots for each language group against a spaced sample of all candidates taking this form of the test in October 1969. A measure of the deviation of each item from the central tendency of the plot was developed, expressing the degree to which the item was especially difficult--or easy--for a particular language group relative to the other items. Distributions of these measures are given for each of the five parts of TOEFL, for each of the six language groups.

PATTERNS OF TEST AND ITEM DIFFICULTY FOR SIX FOREIGN LANGUAGE GROUPS ON THE TEST OF ENGLISH AS A FOREIGN LANGUAGE

William H. Angoff and Amiel T. Sharon
Educational Testing Service

Recent years have seen a significant growth in the numbers of foreign students enrolled in American institutions of higher education. The Institute of International Education (1969) reports that the number of foreign students enrolled in American colleges and universities--121,362 by their count--is more than twice that of a decade earlier. Clearly, with this number of students for whom the English language represents a major stumbling block, and with possible further growth in this number, the need for special language training in English is a serious and pressing problem.

Although many universities currently offer programs to aid and advise foreign students in developing sufficient English language skills to cope with the American culture and university curriculums, much research has yet to be done to make these programs more effective. Many linguists have long felt that foreign students are considerably aided in their English language instruction when special emphasis is given to the specific English language deficiencies and patterns of errors characteristic of students of their own language group.

From the data now available (ETS, 1970) it is clear that the foreign candidates to American colleges who come from different native language groups have different levels of English proficiency.[1] What has not yet been made entirely clear is the extent to which candidates of different language backgrounds have more difficulty with some English language skills

than with others. Some information of this sort is available. King (1968), for example, reports that students applying to U. S. colleges from Afghanistan do much better in listening comprehension than in reading and writing in comparison with applicants from other countries. However, far more information is needed on this topic of differential patterns of linguistic ability, not only in relation to the general skills like listening, reading and writing, but at the more specific and detailed item level. It is to both levels of skill that the present study is directed.

Instruments

The test used in this study to assess the English proficiency of foreign candidates was the Test of English as a Foreign Language (TOEFL). This test is administered by Educational Testing Service domestically and in foreign testing centers throughout the world, and foreign applicants to American institutions of higher education can take the examination in their own native countries. The test consists of five parts, each part measuring an English language skill thought to be important to the foreign student: (1) Listening Comprehension, (2) English Structure, (3) Vocabulary, (4) Reading Comprehension, and (5) Writing Ability. A total score, equal to twice the sum of the scaled scores on the five parts, is also routinely reported for each applicant.

Analysis of the Sections of TOEFL

Method

All members of six native language groups of applicants who took TOEFL in October 1969 for admission to American institutions were chosen as the

subjects for this study. The languages were selected for study on the basis of two considerations: (1) the languages were to be as diverse as possible, representing different branches of the Indo-European family and several different non-Indo-European language groups; (2) they were to be represented by a sufficiently large number of applicants taking TOEFL. The languages selected, the number of subjects within each language group, and each group's test performance (mean and standard deviation) are given in Table 1. In order to provide a general baseline for evaluating the performance of the

---

Insert Table 1 about here

---

candidates in each language group, corresponding statistics are also presented in Table 1 for a "General" group, a sample of 1,000 cases drawn at random from the general candidate group of 14,045 tested at the October 1969 regular administration of TOEFL and used for a routine-type analysis of the form of the test given at that administration.

The data collected in this study were subjected to a series of analyses. The first analysis was a general one, and sought to determine whether the five sections of the test retain the same order of difficulty across all six language groups (exclusive of the "General" group) or whether there was a significant Section x Group interaction--whether some sections of the test were more difficult for some groups than for others relative to other sections of the test. For this purpose a two-factor analysis of variance design was used, as described in Figure 1. The first

---

Insert Figure 1 about here

---

factor was the Sections of the test, which was considered fixed. The second factor, also considered fixed, was Groups; and since an individual could belong to only one of these groups, the category, Individuals, which was considered random, was nested within Groups.

The linear model for the design took the following form:

$$E(X_{jkp}) = \mu + K_k + P_p + J_{j(k)} + KP_{kp} + JP_{jp(k)} \quad ,$$

where

X_{jkp} = the score of the j^{th} person in the k^{th} group on the p^{th} section of the test.

μ = the grand mean.

K_k = the effect of the k^{th} group.

P_p = the effect of the p^{th} section of the test.

$J_{j(k)}$ = the effect of the j^{th} person in the k^{th} group.

KP_{kp} = the interaction of sections and groups.

$JP_{jp(k)}$ = the interaction of sections and persons within groups.

It is also noted that: $j = 1...N_k$; $k = 1...s$; and $p = 1...r$.

The foregoing design permitted the testing of 3 hypotheses:

1. $\sigma^2_k = 0$. There are no differences among the six language groups in the mean scores across the sections of TOEFL (i.e., on total score).

2. $\sigma^2_p = 0$. There are no differences among the five sections of TOEFL for all six language groups combined.

3. $\sigma^2_{kp} = 0$. There is no interaction between the six language groups and the five sections of TOEFL; i.e., there are no significant differences among the groups with respect to their patterns of performance on the five sections of the test.

Table 2 gives a summary of the analysis of variance design.

Insert Table 2 about here

Results

Table 3 gives a summary of the analyses of variance of the parts of

---

Insert Table 3 about here

TOEFL. As would be expected, with as many as 30,600 observations (6,120 individuals in the analysis and five observations per individual), all the testable F-ratios were highly significant.

The fact that the main effect due to Sections was significant simply indicates that for the combined group of 6,120 candidates involved in this study the means on the five sections of TOEFL were not equal. However, as the "Percentage of Total Estimated Variance" column indicates, this source of variance is the smallest of all.

The statistical significance of the effect due to Groups indicates that the various language groups in the study were unequally proficient in the abilities measured by the test. Reference to Table 1 indicates that the native speakers of German were by far the highest performers on all sections of TOEFL, followed in order (with some inconsistencies from one section to another) by speakers of Chinese, Japanese, Gujarati, Arabic, and Spanish. Although the language groups in this study were not intended as representative samples of all TOEFL candidates whose native languages are German, Spanish, Arabic, etc., the fact that the relative order of their performance

is quite similar to that reported in the Interpretive Booklet for TOEFL (ETS, 1970) lends the present study some generality in interpretation.

The principal finding in this analysis is the fact that the Sections x Groups interaction was significant, indicating that the profiles of performance on the five parts of TOEFL are different for the six groups of candidates. This may be seen graphically in Figure 2 and may also be

---

Insert Figure 2 about here

---

observed for any two of the six groups by plotting the pairs of means on the five tests for the two groups, one group on the abscissa, the other on the ordinate. Zero interaction would be represented by a succession of five points all falling on the straight line extending from lower left to upper right. The fact that these points do not fall in a straight line, but show substantial departures from the straight line, illustrates the Section x Group interaction.

The last two columns of Table 3 yield the further information that the components of variance used in this analysis as error--the source attributable to Individuals within Groups and the source attributable to the interaction, Section x Individuals within Groups--represent the largest proportion, almost 81%, of the total of the variance estimates. From that point of view the interaction, Sections x Groups, is not at all sizeable. However, this still does not deny the fact that the Section x Group interaction is itself highly significant, representing a large component of the nonerror variance in the entire analysis. Whatever the proportion of the total of the variance estimates is represented by the Section x Group interaction, it is clear that

the five sections of TOEFL are quite different in difficulty for the six language groups, and different in different ways.

Because the sample sizes varied so considerably from one language group to another--from 116 for the German group to 2853 for the Chinese group--there remained some doubt about the adequacy of the analysis of variance to deal with the Section x Group interaction, in spite of the fact that the size of that F-ratio (236.3 for df = 20; 24,456) very likely precluded the possibility that its statistical significance was in question. Nevertheless, as a check on that possibility, a one-way multivariate analysis of variance was also carried out, testing for the significance of the differences of the five-variable centroids among the six language groups. The results of this analysis confirmed the finding that the Section x Group interaction was significant well beyond the .01 level.[2]

Analysis of Items within Sections of TOEFL

If scores on TOEFL are to be used successfully in diagnosis and remedial training, then it would seem to be important for the teacher of English as a foreign language to familiarize himself with the special English language difficulties of students in each of the major language groups in order to focus the training program to their special linguistic needs. The analyses that are reported below are intended to support this effort by providing a means for studying the interaction of English language proficiencies and native language groups at the item level.

Method

For this purpose, analyses, section by section, were carried out, parallel to the general analysis described above. As before, a two-factor

analysis of variance was carried out as described in Figure 1--except that here "Items" replace "Sections." As before, Groups was taken to be a fixed effect, and Individuals within Groups was random; Items, however, unlike Sections above, was taken to be random. All sections were analyzed in identical fashion, according to a linear model of the following form:

$$E(X_{ijk}) = \mu + I_i + K_k + J_{j(k)} + IK_{ik} + IJ_{ij(k)} \quad ,$$

where

X_{ijk} = the score of the j^{th} person in the k^{th} group on the i^{th} item.

μ = the grand mean.

I_i = the effect of the i^{th} item.

K_k = the effect of the k^{th} group.

$J_{j(k)}$ = the effect of the j^{th} person in the k^{th} group.

IK_{ik} = the interaction of items and groups.

$IJ_{ij(k)}$ = the interaction of items and persons within groups.

It is also noted that: $i = 1...n$; $j = 1...N_k$; $k = 1...s$.

Unlike the earlier design for the analysis of the parts of TOEFL, the foregoing analysis permitted the testing of four hypotheses:

1. $\sigma^2_k = 0$. There are no differences among the six language groups on the section of TOEFL under consideration.

2. $\sigma^2_{j(k)} = 0$. There are no differences among the individuals within the groups on the section of TOEFL under consideration.

3. $\sigma^2_i = 0$. There are no differences in the item difficulties in the section of TOEFL under consideration.

4. $\sigma^2_{ik} = 0$. There is no interaction among the six language groups and the items of the section of TOEFL under consideration; i.e., the rank order of item difficulties (on the particular section of TOEFL) is essentially the same for all six language groups.

Table 4 gives a summary of the analysis of variance design for the items in any one of the five sections of TOEFL. Tables 5-9 give the summaries

---

Insert Tables 4-9 about here

---

of the analyses of variance for the items in Sections 1-5, respectively.

Results

The results shown in Tables 5-9 are quite consistent with one another. In every one of these tables the variances attributable to Groups, to Items, and to Items x Groups are all significant far beyond the one per cent level. The fact that variance due to Items is significant comes as no surprise, since it was intended in the construction of each section to write items that would represent a range of item difficulties. Nor is it surprising that the variance due to Groups is significant. It has already been observed here, in Table 1 and in Table 3, that the six language groups perform differently on the different parts of TOEFL.

As in Table 3, the findings of principal interest in Tables 5, 6, 7, 8, and 9 are that the Items x Groups interaction is significant, giving evidence that even within the five sections of TOEFL the patterns of performance of the six language groups are systematically different. Apparently the items within the sections of the test are measuring sufficiently different

aspects of English language proficiency that they are not uniformly more difficult for one language group than another.

The matter of unequal sample sizes across the different language groups represented the same kind of problem in the analyses of the Item x Group interactions as in the analysis of the Section x Group interaction reported earlier in this paper. However, because the results of the Section x Group analysis were so clearly confirmed by the multivariate analysis of variance, it was felt that, in view of the nature of the results of the Item x Group analyses, a similar confirmation was not necessary here.

Intergroup Comparison of Item Difficulties

The second phase of the present study involves a comparison among the language groups with respect to their performance on the items in each of the five parts of TOEFL. Since it is the purpose of this paper to describe a technique of analysis as much as it is to describe the results of that analysis, it will be useful to devote some space to that technique here.

Method

For each item a percent-pass figure (p-value) was computed, based on the performance of the candidates in each of the six language groups. Each p-value was then converted, by means of the normal probability tables, to a normal deviate, z, and further converted to "delta" by means of the transformation, $\Delta = 4z + 13$. In the comparison of any two groups of examinees (G_x and G_y), the technique involves making a plot of the points, Δ_x versus Δ_y, one point for each of the items under consideration. The plot of points normally falls in an elliptical pattern extending from lower left

to upper right, and if the two groups studied are drawn from the same type of population, then the scatterplot will take the shape of a long, narrow ellipse, often representing a correlation as high as .98 or .99. When the groups are different in level, the points will still fall in a narrow ellipse, displaced vertically or horizontally, depending on which group is the more able one. Even when the groups differ in degree of dispersion, the points will still fall in the same type of ellipse, but in this case the ellipse will be tilted at an angle more or less steeply than 45^{o}, depending on which group is the more dispersed. However, when the groups are different in <u>type</u>, the item difficulties will not fall in the same rank order for the two groups and the correlation represented by the points will be lower. Items falling at some distance from the elliptical swarm and contributing to the item x group interaction are items that are especially easier for one group than for the other, <u>relative to</u> the other items. Such items are deserving of special study. In the context of items like those in TOEFL, they may reveal detailed differences between the groups and may therefore lead to hypotheses regarding underlying differences in the two languages as reflected in the comparative abilities of the two language groups to deal with the nuances of English.

In the present study, with six language groups under consideration, the comparison of each group with every other group would have required the preparation of 15 delta plots for each of the five parts of TOEFL--a total of 75 plots--far too many to present in an article of limited length. In order to reduce sharply the number of such comparisons it was decided to compare each of the six language groups with a single "standard," or General, group, drawn at random from the entire candidate population of about 14,000

taking this form of the test at its first operational administration in October 1969 and representing candidates from many different language groups.[3]

Although the decision to compare each group with a single standard (General) group did in fact reduce the number of comparative Δ-plots, there were still too many to present in detail, and, accordingly, a method was sought to summarize the significant features of the comparisons that would be made possible by those plots. For each elliptical plot of items the equation of the major axis of the ellipse was determined, and the perpendicular distance (D_i) from each point to that line was calculated. The variance of the distribution of these distances represents the item x group interaction and the distribution of these distances makes it possible to highlight those items that deviate sharply from the swarm. The equation used for the major axis of the ellipse was $Y = AX + B$, where

$$A = \frac{(\sigma_y^2 - \sigma_x^2) \pm \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4r_{xy}^2 \sigma_x^2 \sigma_y^2}}{2r_{xy} \sigma_x \sigma_y}$$

and

$$B = M_y - AM_x \quad .$$

(It is recalled that the variables, x and y, are, respectively, the delta values for the two groups under consideration.) The formula for the perpendicular distance, D_i, of each point, i, in the plot to the line is given as:

$$D_i = \frac{AX_i + B - Y_i}{\sqrt{A^2 + 1}} \quad .$$

Results

The distributions of the D-values in the plots for each language group versus the General candidate group are given in Tables 10-14, each

---

Insert Tables 10-14 about here

---

table corresponding to one of the five parts of TOEFL. In each distribution of the D-values in the plot of a particular language group versus the General group, the items with positive D-values are more difficult for the particular language group, relative to the other items in the elliptical swarm; the items with negative D-values are those that are more difficult for the General group, relative to the other items. This is not to say necessarily that the items on the positive or negative side of the zero point are more difficult for one group than for another. It is only that they are relatively more difficult for one group or the other in comparison with the other items. For example, in the very first distribution of Table 10, giving the D-values in the plot of deltas for German-speaking candidates versus the candidates in the General group, only one item of the 49 plotted was more difficult for the German speakers than for the General candidate group.

In addition to the number of items tabulated in each distribution (which would equal the number of items in the test except when, for example, an item is answered correctly by everyone, or by no one, in the group, and a Δ -value and D-value cannot be calculated for it), the mean D-value is given (always zero), as well as the standard deviation of D-values and the correlation, r_{ig}, represented by the item plot. Both the standard deviation of D-values and the correlations between delta-values describe the extent

of the item x group interaction. In one respect the correlation between deltas is the preferable index, since it is a pure number, unaffected, as is σ_D, by the difference in the standard deviations of deltas for the two groups. The value of r_{ig} can be used as an index of the similarity in the types of errors in English made by the two groups under consideration; and to the extent that their errors in English are indicative of the structure of their own language and the test items in English are sensitive to that structure, the value of r_{ig} can be used as an index of the similarities in those structures. As expected, the different parts of TOEFL will show the different language groups to be similar in different degrees. It is clear that the Listening Comprehension section of TOEFL yields interlanguage group plots that have the highest r_{ig} values ($\overline{r}_{ig} = .914$) of all the sections of the test, and in that sense the Listening Comprehension section represents perhaps the most homogeneous of all the sections of TOEFL for these six language groups. The Reading Comprehension section also yields plots with high r_{ig} values ($\overline{r}_{ig} = .894$), possibly because the contextual cues in a reading passage tend to equalize the influence of native language on comprehension. These higher correlations (and lower σ_D-values) for the Listening and Reading Comprehension sections reflect a result already observed in Tables 5-9, that the Item x Group interactions for these sections are smaller than for the other three sections of the test. The lowest value of $\overline{r}_{ig}$ (.782) was found in the Writing Ability section. Therefore it would be assumed that the Writing Ability section--possibly in addition to English Structure ($\overline{r}_{ig} = .826$) and Vocabulary ($\overline{r}_{ig} = .824$)--might be, generally speaking, the most revealing sections of TOEFL for purposes of linguistic analysis. At a more specific level it would appear that the English Structure section is most revealing for the Chinese ($r_{ig} = .742$) and Gujarati ($r_{ig} = .752$)

groups; the Vocabulary section is most useful for the Spanish (r_{ig} = .773) and possibly Chinese (.771) groups; and the Writing Ability section most revealing for the German (r_{ig} = .689) and Spanish (r_{ig} = .674) groups.

Summary

Two-factor analyses of variance with multiple measurements on one factor were conducted among the five sections of TOEFL and six language groups, and also among the items of each section. All sources of variance were found to be significant beyond the one per cent level. Of particular interest was the fact that the Test x Group interactions were highly significant and that the Item x Group interactions within test were also highly significant, indicating that some tests and some items were relatively more difficult for some groups than for others.

Further displays of the Item x Group interactions were made by analyzing the item difficulty plots for each language group against a spaced sample of all candidates taking this form of the test at its first formal administration. A measure of the deviation of each item from the central tendency of the plot was developed, expressing the degree to which the item was especially difficult--or easy--for a particular language group relative to the other items. Distributions of these measures are given for each of the five parts of TOEFL, for each of the six language groups.

Although this paper makes no attempt to analyze the Section x Group and Item x Group interactions in terms of linguistic considerations, it does attempt to evaluate the statistical justification for making such an analysis. It is then left to any investigator with competence in linguistic analysis to make such an analysis, if he wishes to, by examining the items in detail along with the statistical data on the items that will be made available to him on request.

References

Educational Testing Service. Test of English as a Foreign Language, Interpretive Information. Princeton, N. J.: ETS, 1970.

Institute of International Education. Open Doors 1969. New York: IIE, 1969.

King, R. A. The Test of English as a Foreign Language in Afghanistan. Kabul, Afghanistan: USAID/TCCU, 1968.

Footnotes

[1]The data referred to here are the summaries of performance of candidates who take the Test of English as a Foreign Language (TOEFL) as part of their requirements for admission to American universities. It is entirely likely that the differences in level as shown in the TOEFL data only reflect the different self-selective factors operating on these students in their native countries, but not the levels of performance of all English speakers in those countries.

[2]The authors are indebted to F. Reid Creech for his invaluable assistance in carrying out the multivariate analysis of variance for these data.

[3]The authors wish to express their appreciation to Dr. John B. Carroll for suggesting this solution.

Table 1

Scaled Score Means and Standard Deviations on TOEFL
for Six Language Groups and for the General Group (Test Analysis Sample)

Language Group	No. of Cases	Listening Comprehension		English Structure		Vocabulary		Reading Comprehension		Writing Ability		Total Score	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
German	116	61.1	7.3	55.0	6.2	55.7	7.5	59.2	4.4	55.5	8.5	573	57
Spanish	1411	45.0	14.2	42.2	10.5	44.4	10.1	46.9	9.0	40.6	11.0	438	102
Arabic	286	47.9	10.9	45.3	8.9	40.6	10.8	46.1	7.8	44.7	9.8	449	87
Chinese	2853	46.7	8.9	53.1	6.0	49.1	8.5	48.4	6.2	50.5	7.0	496	58
Japanese	740	46.7	9.6	50.2	7.3	46.7	9.7	49.4	7.4	48.3	8.9	483	74
Gujarati	714	42.0	10.3	43.5	9.7	46.3	11.1	48.9	8.4	47.2	9.7	456	87
General	1000	49.0	11.0	50.3	8.6	49.4	10.6	50.5	8.0	49.8	9.8	498	82

Table 2

Design of the Analysis of the Sections of TOEFL

Source of Variation	Degrees of Freedom	E(MS)
Between Individuals	$\sum_{k}^{s} N_k - 1$	
Groups [k]	$s - 1$	$\bar{N}r\sigma^2_k + r\sigma^2_{j(k)}$
Indivs within Groups [j(k)]	$\sum_{k}^{s} N_k - s$	$r\sigma^2_{j(k)}$
Within Individuals	$(r - 1)\sum_{k}^{s} N_k$	
Sections [p]	$r - 1$	$\bar{N}s\sigma^2_p + \sigma^2_{jp(k)}$
Sections x Groups [kp]	$(r - 1)(s - 1)$	$\bar{N}\sigma^2_{kp} + \sigma^2_{jp(k)}$
Sections x Indivs wn Groups [jp(k)]	$(r - 1)(\sum_{k}^{s} N_k - s)$	$\sigma^2_{jp(k)}$

Table 3

Summary of Analysis of Variance for Sections of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	35,589.74				
Groups	5	4,093.94	818.79	158.9*	.1903	10.6
Indivs within Groups	6,114	31,495.80	5.1514		1.0303	57.2
Within Individuals	24,480	13,195.22				
Sections of the Test	4	808.40	202.10	476.1*	.0393	2.2
Sections x Groups	20	2,005.74	100.29	236.3*	.1168	6.5
Sects x Indivs wn Groups	24,456	10,381.08	0.42448		.4245	23.6
Total	30,599	48,784.96				

*$p < .01$

Table 4

Design of the Analysis of the Items in Any Section of TOEFL

Source of Variation	Degrees of Freedom	E(MS)
Between Individuals	$\sum_{k}^{s} N_k - 1$	
Group [k]	$s - 1$	$n\bar{N}\sigma^2_k + \bar{N}\sigma^2_{ik} + n\sigma^2_{j(k)} + \sigma^2_{ij(k)}$
Indivs within Groups [j(k)]	$\sum_{k}^{s} N_k - s$	$n\sigma^2_{j(k)} + \sigma^2_{ij(k)}$
Within Individuals	$(n - 1) \sum_{k}^{s} N_k$	
Items [i]	$n - 1$	$s\bar{N}\sigma^2_i + \sigma^2_{ij(k)}$
Items x Groups [i x k]	$(n - 1)(s - 1)$	$\bar{N}\sigma^2_{ik} + \sigma^2_{ij(k)}$
Items x Indivs wn Groups [ij(k)]	$(n - 1)(\sum_{k}^{s} N_k - s)$	$\sigma^2_{ij(k)}$

Table 5

Summary of Analysis of Variance
of the Items in the Listening Comprehension Section of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	12,042.27				
Groups	5	688.99	137.80	19.9*†	.0031	1.2
Indivs within Groups	6,114	11,353.28	1.8569	10.3*	.0334	13.1
Within Individuals	299,880	64,007.70				
Items	49	8,662.04	176.78	979.7*	.0341	13.4
Items x Groups	245	1,286.43	5.2507	29.1*	.0040	1.6
Items x Indivs wn Groups	299,586	54,059.23	.18044		.1804	70.7
Total	305,999	76,049.97				

*$p < .01$

†Quasi F ratio

Table 6

Summary of Analysis of Variance
of the Items in the English Structure Section of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	10,481.93				
Groups	5	2,835.88	567.18	48.3*†	.0162	6.5
Indivs within Groups	6,114	7,646.05	1.2506	7.3*	.0270	10.8
Within Individuals	238,680	47,686.95				
Items	39	5,037.10	129.16	759.1*	.0251	10.0
Items x Groups	195	2,077.35	10.653	62.6*	.0123	4.9
Items x Indivs wn Groups	238,446	40,572.50	.17015		.1702	67.9
Total	244,799	58,168.88				

*$p < .01$

†Quasi F ratio

Table 7

Summary of Analysis of Variance
of the Items in the Vocabulary Section of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	8,834.56				
Groups	5	647.88	129.58	6.9*†	.0032	1.2
Indivs within Groups	6,114	8,186.67	1.3390	7.6*	.0291	11.2
Within Individuals	238,680	51,524.38				
Items	39	6,262.07	160.57	915.8*	.0313	12.1
Items x Groups	195	3,454.16	17.714	101.0*	.0205	7.9
Items x Indivs wn Groups	238,446	41,808.14	.17534		.1753	67.6
Total	244,799	60,358.93				

*$p < .01$

†Quasi F ratio

Table 8

Summary of Analysis of Variance
of the Items in the Reading Comprehension Section of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	8,487.49				
Groups	5	458.82	91.763	18.7*†	.0034	1.3
Indivs within Groups	6,114	8,028.67	1.3132	6.8*	.0373	14.8
Within Individuals	177,480	37,007.90				
Items	29	2,092.20	72.145	372.2*	.0140	5.5
Items x Groups	145	549.44	3.7892	19.5*	.0042	1.7
Items x Indivs wn Groups	177,306	34,366.26	.19382		.1938	76.7
Total	183,599	45,495.39				

*$p < .01$

†Quasi F ratio

Table 9

Summary of Analysis of Variance
of the Items in the Writing Ability Section of TOEFL

Source of Variation	df	SS	MS	F	Est. of Variance	Per Cent of Total
Between Individuals	6,119	8,130.33				
Groups	5	1,468.12	293.62	26.1*†	.0083	3.2
Indivs within Groups	6,114	6,662.21	1.0897	5.7*	.0225	8.8
Within Individuals	238,680	52,084.08				
Items	39	4,697.01	120.44	633.0*	.0234	9.1
Items x Groups	195	2,018.53	10.351	54.4*	.0119	4.6
Items x Indivs wn Groups	238,446	45,368.54	.19027		.1903	74.2
Total	244,799	60,214.40				

*$p < .01$

†Quasi F ratio

Table 10

Distributions of D-Values
for the Listening Comprehension Section

	German	Spanish	Arabic	Chinese	Japanese	Gujarati
1.6 - 1.7	4					
1.4 - 1.5	0			1		1
1.2 - 1.3	1		3	0		0
1.0 - 1.1	3	1	2	1	1	2
0.8 - 0.9	2	1	1	1	1	3
0.6 - 0.7	1	4	3	4	5	1
0.4 - 0.5	1	5	2	7	4	10
0.2 - 0.3	6	6	7	5	7	3
0.0 - 0.1	3	9	7	8	8	3
-0.2 - -0.1	7	11	10	9	13	10
-0.4 - -0.3	6	7	6	5	5	7
-0.6 - -0.5	5	2	3	3	1	3
-0.8 - -0.7	4	3	2	2	3	4
-1.0 - -0.9	2	1	1	3	1	1
-1.2 - -1.1	3		2	1	0	2
-1.4 - -1.3	1		1		1	
No. of Items*	50					
Mean	0.0					
Std. Dev.	0.780	0.419	0.581	0.536	0.464	0.582
r_{ig}	.857	.940	.906	.931	.949	.902

*D-value for one item in the German group was not tabulated.

Table 11

Distributions of D-Values
for the English Structure Section

	German	Spanish	Arabic	Chinese	Japanese	Gujarati
2.0 - 2.1	1					1
1.8 - 1.9	1					0
1.6 - 1.7	0			2		0
1.4 - 1.5	0			1	2	2
1.2 - 1.3	0		1	1	0	1
1.0 - 1.1	1	1	2	4	3	2
0.8 - 0.9	3	1	4	1	3	2
0.6 - 0.7	2	5	2	1	4	5
0.4 - 0.5	6	3	4	3	3	1
0.2 - 0.3	6	6	7	6	4	5
0.0 - 0.1	2	9	4	3	3	1
-0.2 - -0.1	3	4	2	2	4	5
-0.4 - -0.3	4	4	3	2	3	1
-0.6 - -0.5	5	2	3	6	2	7
-0.8 - -0.7	1	1	3	1	3	0
-1.0 - -0.9	0	2	2	2	0	1
-1.2 - -1.1	1	1	1	0	3	0
-1.4 - -1.3	3	1	1	3	0	4
-1.6 - -1.5	0		0	0	1	2
-1.8 - -1.7	1		0	1	1	
-2.0 - -1.9			1	0	1	
-2.2 - -2.1				1		
No. of Items	40					
Mean	0.0					
Std. Dev.	0.789	0.522	0.716	0.912	0.831	0.896
r_{ig}	.862	.922	.860	.742	.818	.752

Table 12

Distributions of D-Values
for the Vocabulary Section

	German	Spanish	Arabic	Chinese	Japanese	Gujarati
1.8 - 1.9	1	2				
1.6 - 1.7	0	2		1		
1.4 - 1.5	3	3		1		1
1.2 - 1.3	3	1		2	1	0
1.0 - 1.1	1	3	2	2	0	2
0.8 - 0.9	0	1	1	1	3	3
0.6 - 0.7	1	0	8	5	3	2
0.4 - 0.5	4	1	3	3	5	5
0.2 - 0.3	6	2	4	5	9	4
0.0 - 0.1	2	3	7	3	2	7
-0.2 - -0.1	4	3	5	1	2	2
-0.4 - -0.3	2	6	2	2	6	6
-0.6 - -0.5	2	1	1	2	3	2
-0.8 - -0.7	2	3	2	4	3	1
-1.0 - -0.9	3	3	1	5	1	2
-1.2 - -1.1	4	2	1	2	1	0
-1.4 - -1.3	0	1	1	0	0	1
-1.6 - -1.5	1	1	1	0	0	2
-1.8 - -1.7	1	1	1	0	0	
-2.0 - -1.9		0		0	1	
-2.2 - -2.1		0		0		
-2.4 - -2.3		0		0		
-2.6 - -2.5		1		1		
No. of Items			40			
Mean			0.0			
Std. Dev.	0.897	1.091	0.676	0.893	0.640	0.676
r_{ig}	.821	.733	.858	.771	.881	.878

Table 13

Distributions of D-Values
for the Reading Comprehension Section

	German	Spanish	Arabic	Chinese	Japanese	Gujarati
1.4 - 1.5	1					
1.2 - 1.3	0	1				1
1.0 - 1.1	1	1	1			1
0.8 - 0.9	2	1	1		3	0
0.6 - 0.7	3	3	2		0	0
0.4 - 0.5	2	2	2	6	6	4
0.2 - 0.3	3	6	3	7	3	6
0.0 - 0.1	0	2	3	3	5	2
-0.2 - -0.1	7	2	9	7	4	6
-0.4 - -0.3	2	5	6	3	3	4
-0.6 - -0.5	2	3	2	3	3	4
-0.8 - -0.7	3	3	1	1	3	2
-1.0 - -0.9	3	0				
-1.2 - -1.1		1				
-1.4 - -1.3						
-1.6 - -1.5						
No. of Items*	30					
Mean	0.0					
Std. Dev.	0.638	0.596	0.416	0.342	0.457	0.479
r_{ig}	.825	.833	.925	.957	.918	.905

*D-value for one item in the German group was not tabulated.

Table 14

Distributions of D-Values
for the Writing Ability Section

	German	Spanish	Arabic	Chinese	Japanese	Gujarati
2.6 - 2.7	1					
2.4 - 2.5	0					
2.2 - 2.3	0					
2.0 - 2.1	1					
1.8 - 1.9	0					
1.6 - 1.7	0		1			
1.4 - 1.5	3	3	0	1	1	1
1.2 - 1.3	0	3	2	0	1	0
1.0 - 1.1	2	1	0	3	2	4
0.8 - 0.9	1	1	0	1	0	0
0.6 - 0.7	4	2	2	6	3	3
0.4 - 0.5	2	4	7	2	2	4
0.2 - 0.3	2	3	7	7	6	4
0.0 - 0.1	2	4	2	3	5	5
-0.2 - -0.1	6	1	4	3	2	5
-0.4 - -0.3	1	2	4	2	12	3
-0.6 - -0.5	2	8	4	5	4	3
-0.8 - -0.7	4	4	2	1	0	6
-1.0 - -0.9	3	1	4	3	0	1
-1.2 - -1.1	3	0	1	1	1	0
-1.4 - -1.3	2	1		1	1	0
-1.6 - -1.5	1	1		0		1
-1.8 - -1.7		1		0		
-2.0 - -1.9				0		
-2.2 - -2.1				1		
No. of Items			40			
Mean			0.0			
Std. Dev.	0.961	0.832	0.648	0.736	0.585	0.656
r_{ig}	.689	.674	.838	.812	.868	.810

Figure Captions

Fig. 1. Diagrammatic representation of the data for the analysis of variance.

Fig. 2. Profiles of scores for the six language groups on the sections of TOEFL.

		Sections			
	Individuals	1	2	.....p.....	r
Group 1	1 2 ⋮ N_1				
Group 2	1 2 ⋮ N_2				
	⋮				
Group k	1 2 ⋮ N_k				
	⋮				
Group s	1 2 ⋮ N_s				

Fig. 1.

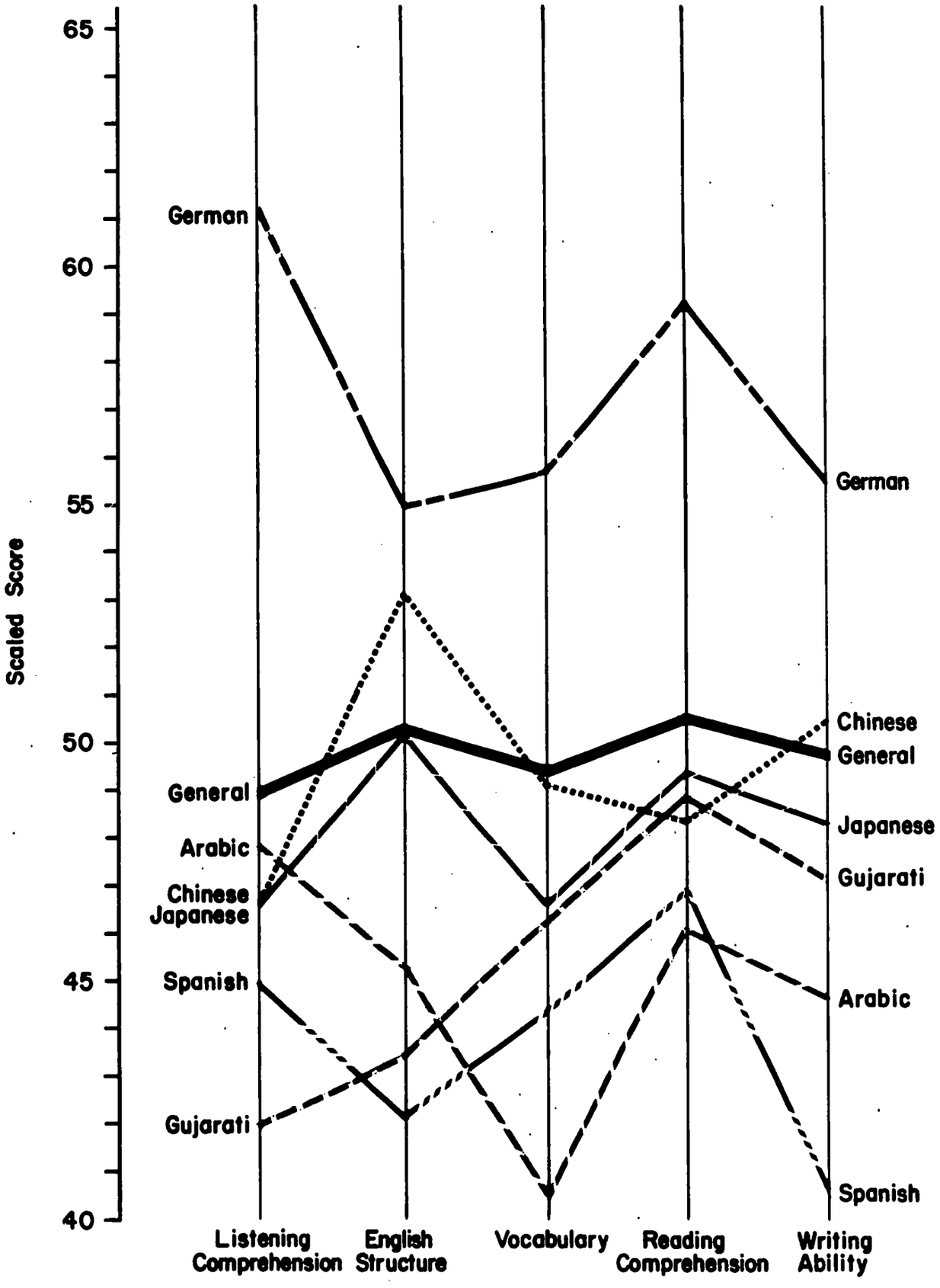
65
60
55
50
45
40
Scaled Score
German
General
Arabic
Chinese
Japanese
Spanish
Gujarati
German
Chinese
General
Japanese
Gujarati
Arabic
Spanish
Listening Comprehension
English Structure
Vocabulary
Reading Comprehension
Writing Ability

Fig. 2.