DOCUMENT RESUME

ED 385 599 TM 024 047

AUTHOR Boldt, R. F.; And Others

TITLE Distribution of ACTFL Ratings by TOEFL Score Ranges.

INSTITUTION Educational Testing Service, Princeton, N.J.

REPORT NO ETS-RR-92-59; TOEFL-RR-41

PUB DATE Nov 92

NOTE 57p.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS *College Students; Correlation; *English (Second Language); Higher Education; *Language Teachers; Performance; Rating Scales; *Scores; Second Languages; *Statistical Distributions; Test Results; Test Use

IDENTIFIERS *American Council on the Teaching of Foreign Langs; *Test of English as a Foreign Language

ABSTRACT

The purpose of this study was to align verbal descriptions of test takers' language performance with distributions of the numerical scores they received on the three sections (Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary) of the Test of English as a Foreign Language (TOEFL). The descriptors of the American Council of Teaching of Foreign Languages (ACTFL) were used as anchors for the TOEFL scores. Eighty-four English-as-a-Second-Language instructors rated the listening, reading, and writing proficiency of students (from 60 to 150 at each of 7 institutions) using the ACTFL descriptors. The ratees then took the TOEFL. Students' ACTFL ratings were quantified and cross-tabulated with TOEFL section scores. Although there was no one-to-one correspondence between the TOEFL score level and the ACTFL rating level, the many substantial correlations between test scores and ratings provided evidence that the ACTFL ratings and TOEFL scores tap similar underlying skills. Distributions of the ratings at levels of the TOEFL scores were developed to help interpret TOEFL scores in terms of language performance. Seven tables and one figure illustrate the analyses. Three appendixes give ACTFL guidelines, the rating booklet, and percentile distributions for ACTFL ratings. (Contains 13 references.) (Author/SLD)

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

ED 385 599

TEST OF ENGLISH AS A FOREIGN LANGUAGE

# Research Reports

REPORT 41
November 1992

## Distributions of ACTFL Ratings by TOEFL Score Ranges

R.F. Boldt
D. Larsen-Freeman
M.S. Reed
R.G. Courtney

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
□ Minor changes have been made to improve reproduction quality
▪ Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service

BEST COPY AVAILABLE

TM024041

# DISTRIBUTIONS OF ACTFL RATINGS BY TOEFL SCORE RANGES

R. F. Boldt, D. Larsen-Freeman, M. S. Reed, and R. G. Courtney

Educational Testing Service
Princeton, New Jersey

RR-92-59

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

Copyright © 1992 by Educational Testing Service. All right reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both US and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service.

ERIC
Full Text Provided by ERIC

## Abstract

The purpose of this study was to align verbal descriptions of test takers' language performance with distributions of the numerical scores they received on the three sections (Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary) of the Test of English as a Foreign Language (TOEFL).

The study used the American Council of Teaching Foreign Languages (ACTFL) descriptors as anchors for the TOEFL scores. The ACTFL Proficiency Guidelines, from which the descriptors were taken, are widely used.

Qualified English as a Second Language (ESL) instructors rated their students' listening, reading, and writing proficiency using the ACTFL descriptors. Very shortly thereafter, the ratees took the TOEFL. Students' ACTFL ratings were quantified and cross-tabulated with TOEFL section scores.

Results were analyzed as follows: First, the need to adjust for rater difficulty was investigated. Second, typical descriptive data were developed. Third, several schemes were used to quantify the verbal descriptors from the ACTFL Guidelines. Fourth, the ratings and test scores were correlated. Fifth, the issue of the differential validity of the measures of speaking, listening, and reading was examined. Sixth, several estimates of the reliability of the ratings were obtained. Finally, the distributions of ratings at levels of the TOEFL section scores were developed. These latter distributions can be helpful in interpreting TOEFL scores in terms of language performance. Although there was no one-to-one correspondence between the TOEFL score level and the ACTFL rating level, the correlations between ratings and scores were substantial.

ERIC
Full Text Provided by ERIC

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

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

✣ ✣ ✣

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

| | |
|---|---|
| James Dean Brown | University of Hawaii |
| Patricia Dunkel (Chair) | Pennsylvania State University |
| William Grabe | Northern Arizona University |
| Kyle Perkins | Southern Illinois University at Carbondale |
| Elizabeth C. Traugott | Stanford University |
| John Upshur | Concordia University |

## Table of Contents

## Figure

## List of Tables

## Introduction

The purpose of this study was to provide a statistical alignment of narrative descriptors with different score levels, or score ranges, of the three TOEFL sections--Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. The descriptors were to characterize the language-related performance of the examinees at the different score levels.

Systematic assessment of the English language proficiency of those for whom English is not the first language is done for a variety of purposes and can be accomplished using several measures. Two of the most important are multiple-choice testing, as with the Test of English as a Foreign Language (TOEFL), and anchored rating, as with the American Council on the Teaching of Foreign Languages' (ACTFL) rating system (ACTFL, 1986). When used to assess the level of comparable language skills (listening, writing, and reading), these measures should, under the proper circumstances, result in assessments that agree. This study examined relationships between these two types of measures for the skills of listening, writing, and reading.

## Background

### ACTFL Descriptors for TOEFL Scores

Some educators have suggested that TOEFL scores would be more meaningful if ratings were aligned with specific descriptors of language performance, such as those used with the ACTFL rating system. The purpose of the present study was to align such descriptors with different score levels, or score ranges, for the three TOEFL sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. The general plan of the study was to obtain ratings of language proficiency using the ACTFL descriptors followed soon by the administration of the TOEFL. The ratings and scores could then be cross-tabulated to determine the distributions of ratings assigned to examinees who scored in specified TOEFL score ranges.

The Interagency Language Roundtable's language skill level descriptions (ILR, 1985), which are couched in explicitly functional language terms, approximate the descriptors needed here. The ILR guidelines, which are used by approximately 30 U.S. government agencies, include separate guidelines for listening, reading, and writing skills corresponding to the three TOEFL sections. They are intended for use by qualified raters to describe ratees' levels of language proficiency, a purpose that is quite consistent with our own. Their history is described by ETS (1982) and by Lowe (1988).

The American Council on the Teaching of Foreign Languages has developed proficiency guidelines that, as Lowe points out, are closely related to the ILR descriptors but apply to an educational context. Their functional nature is shown by the following examples from the ACTFL (1986) guidelines: (a) "able to understand short, learned utterances, particularly where content strongly supports understanding and speech is clearly audible;" (b) "able to understand parts of texts which are conceptually abstract and linguistically complex;" (c) "material produced consists of recombinations of learned vocabulary and structures into simple sentences on very familiar topics."[1] The functional orientation in an educational context in the ACTFL guidelines was desired for the present project. The listening, writing, and reading scales correspond to the three TOEFL sections and are given in Appendix A.

## Rating Scales and Data Collection

### Revision of the Existing Ratings

At the beginning of this study, we planned to use the ACTFL materials in developing descriptors more suitable for the TOEFL program. We attempted this development because the descriptors would be interpreted by those without specific training. Also, a number of criticisms of the ACTFL ratings have appeared (Bachman and Savignon, 1986; Lantolf and Frawley, 1985; Savignon 1985; Lantolf and Frawley 1988; Bachman 1988; Douglas 1988). It should be noted that many of these comments are given in the context of speaking rather than listening, writing, and reading that are of concern here, and they are based on logic rather than empirical data. In contrast, an empirical study by Dandanoli and Henning (1990) provided support for the construct validity of the ACTFL speaking scale.

We rewrote the descriptors in an attempt to provide more concrete definitions of terms for the receptive skills (e.g. reading and listening), particularly to provide subjectively equal spacing of the proficiency levels described, and to simplify the language. However, when the revisions were reviewed, we felt that the improvements were not great enough to do without the existing ACTFL guidelines, which are well known and widely used. An enhancement of the descriptors might result in sharper relationships but would require a substantially greater level of effort beyond that contemplated here. Also, associating test scores with ratings based on existing guidelines would be useful in its own right.

[1] Example A was taken from the Generic Descriptions for listening at the Novice Mid level; example B is at the Advanced Plus level for reading; example C is at the Intermediate Low level for writing.

Raters

There were 84 raters across seven institutions, who were experts familiar with the language proficiency of the ratees. All but two of the raters had three of the following five qualifications: a master's degree in ESL or Applied Linguistics, experience teaching ESL, several years of teaching experience, proficiency in another language, and experience living overseas. The two exceptions had bachelor's rather than master's degrees and had not lived overseas but were included because their ESL teaching experience was extensive--13 and 18 years, respectively.

All of the raters were affiliated with university intensive ESL programs offering at least 15 hours of class instruction per week. The programs were open to all foreign nationals and to students of varying levels of English proficiency. All programs administered the institutional TOEFL examination.

Participating Institutions

The participating institutions were colleges and universities in the eastern United States from Boston to Miami. Their ESL programs were organized around quarter or semester systems with three to five classes per level including reading, listening, writing, TOEFL preparation, and other classes designed to prepare students for college entry. More than 60 countries were represented among the international student populations. Students were grouped according to proficiency levels; each school had its own system for defining levels. The institutions had enrollments that ranged from 60 to 150 students. There were between 10 and 30 instructors per site; individual students usually had from 3 to 5 instructors.

Rating Procedure

A member of the Educational Testing Service staff visited each participating institution, introduced the study and materials, supplied rating forms, and answered questions. The rating occurred during scheduled site visits. Only one visit per institution was needed for all institutions except one. This institution required two visits because the raters for listening, writing, and reading were not all available on the same day, not because of any problems with the rating procedures.

The site visits occurred within one to fourteen days after the institutional TOEFL administration, with an average of 6.6 days. In all but one case, the ratings were made prior to the test administration. In that one instance, the test occurred the day before the collection of rating data. At no time were the TOEFL scores available to the raters at the time of the rating.

We asked the raters to evaluate the proficiency of as many students as possible, with the restriction that they be familiar with the ratees' proficiency on the skill being rated. Appendix B contains instructions and a sample rating page for reading. 2/

Note that the rating procedure does not depend on exposure of the raters to the ACTFL Guidelines prior to the rating sessions. The raters were not certified; indeed, there is no formal ACTFL certification as listening, writing, or reading raters. Even if there were such a program, there were insufficient resources available to the study to use it. In this study, we relied on the raters' common interpretation descriptors in terms of ratee language behavior.

## Analyses and Results

The analysis focused on several areas, a crucial one being whether rater tendencies should affect other analyses. The investigation of "rater difficulty" is thus the first analysis presented. Other analyses include descriptive statistics, quantification of the ACTFL scale, discriminant validity, reliability, and cross-tabulation of ACTFL ratings with TOEFL scores.

### Adjustment for Rater Difficulty

The ACTFL descriptors rather ambiguously convey the levels of proficiency intended. Their language is somewhat relative without specifying the norm reference; terms such as "comprehends" do not indicate what behavior to examine to determine if comprehension has occurred. Using the ACTFL paragraphs as a gauge of students' proficiency is certainly not like reading a meter. The ratings may reflect raters' tendencies to be severe or lax more than they do the precise quality level of the behavior being evaluated. To explore this possibility, an analysis was conducted in which ratee scores were calculated in three ways, two of which take into account the overall tendency of raters to give low or high ratings, and a third that does not take the rater tendencies into account. A ratee's score calculated using the third method consisted of the average rating assigned regardless of which raters provided the ratings. The three methods were then evaluated in terms of how well they summarized the individual ratings, and how they correlated with TOEFL scores.

One method of correcting ratee scores for rater tendencies is based on assumptions defining the "football correction." The idea of the football correction is a familiar one: If team A and

2/We appreciate Grant Henning's help as a sounding board and advisor with respect to the data collection layout.

team B have a common opponent that team A beats by 4 points and team B by 1 point (A is 3 better than B); and if team B and team C have a common opponent that team B beats by 8 points and team C beats by 6 points (B is 2 better than C), then we expect team A to beat team C by 5 points ((4-1)+(8-6)). Bettors know from sad experience that this sort of transitivity is not precise. However, if a set of teams plays each other many times so that there are many comparisons, an averaging process can be used to rate the teams. We can build a handicapping system this way that will work reasonably well.

In our situation, we chose a number to be assigned to each ratee and a number to be assigned to each rater. The expected rating for a particular ratee by a particular rater is the sum of their two numbers. Then if ratee A and ratee B have a rater in common who assigns a 4 to A and a 1 to B (A is 3 better than B), and if ratee B and ratee C have a common rater who assigns an 8 to B and a 6 to C (B is 2 better than C), we expect a common rater to rate A 5 points better than C ((4-1)+(8-6)). Of course, at the start of the analysis, we do not know the numbers but we can fit them in a least-squares sense if we have enough interconnections. Therefore, the football correction yields a ratee score that is adjusted automatically for the tendency of the rater to give high or low numbers.

As described above, the football correction was adjusted for rater difficulty by adding a rater constant--what we would call an "additive adjustment." But the adjustment could go a step further and adjust for raters' differing tendencies to use the whole scale. We could do this by assigning two numbers to a rater--one to add and one to multiply by. The multiplicative adjustment is essentially a stretching factor used along with the additive adjustment. Again, the ratee scores and rater adjustments can be calculated by least-squares methods--it amounts to extracting a single factor, as in factor analysis, with missing data. We may call this correction the "linear correction" because the ratee's score is translated into an expected rating by a linear transformation appropriate to the particular rater involved.

The two methods of treating raters' tendencies to be severe or lax were evaluated as follows. The variance of the corrected ratings around the uncorrected ratings was computed and subtracted from the variance of the uncorrected ratings. The resulting figure was transformed to a fraction of the variance of the uncorrected ratings; the fraction was subtracted from one and the square root taken. This yielded a figure that is the analog of a correlation in that its square is the fraction of variance accounted for by the correction. In Table 1, it is referred to as "r with Rating." Second, the "corrected" ratings were evaluated in terms of their correlations with TOEFL section scores. To do this, the corrected ratings were correlated with the TOEFL score.

In Table 1, this figure is referred to as "r with TOEFL." Similar methods were used in Table 1 to obtain the entries under "r with Rating" and "r with TOEFL" in the line labeled "Uncorrected."

The results for this analysis are presented in Table 1. It was not possible to present data for all schools, but a sufficient number of cases with multiple ratings were available for an analysis of listening ratings at School C and of all three skills at School E. In this table, notice that, for School C on the listening rating, by moving from the uncorrected to the football and linear corrections, the approximation of the original ratings is improved, with correlations being .78, .92, and .95, respectively. However, in the left-hand column, we see that the correlations with the TOEFL listening scores are quite close: .74, .76, and .76, respectively. Because the number of parameters used is least for the uncorrected method and greatest for the linear correction, the increase across the three rating corrections shown in the correlations with ratings could be due to capitalization on chance. That this may well be the case can be seen from the failure of the correlations with TOEFL to increase. Similar results can be found in the data from School E whose correlations with original ratings are quite large, being in the high 80s and 90s. As with School C, the correlations for School E with TOEFL are lower. For reading at School E, the effect is exaggerated because of the quite small numbers of degrees of freedom, with the correlations with ratings being all in the high 90s but the correlations with TOEFL being .60. We will not consider the adjustment for severity or leniency further because it does not affect the correlations with TOEFL (the "r with TOEFL" correlations are all approximately the same despite the means used to calculate the rating). When several raters evaluate a student, those ratings will be simply averaged.

## Descriptive Statistics

Table 2 presents means, standard deviations, and numbers of ratees and raters for the seven participating institutions as well as a summary section. The ratings were quantified using the "equal interval" scale (see below).

Note that though the ratings ranged from one to 10, the averages varied around six. This occurred because very few ratings of one were noted. Informal inquiries revealed that students who would have received a rating of one were regarded as unready for TOEFL, and, hence, would not have appeared in the study. Note also that institution G had the highest ratings and the lowest TOEFL scores (see paragraph at end of Appendix C).

## Quantification of the ACTFL Scale

Quantification of the ACTFL descriptors was useful for statistical purposes, such as computing descriptive statistics and reliabilities. A question about quantification was whether to regard ACTFL descriptors as defining equal intervals of proficiency. This question was examined by using an equal-interval scale and one other in computing descriptive statistics, part of the discriminant validity analysis, and the reliability analysis. The results using these two scales could then be compared.

The equal-interval scale assigns the scale value of an ACTFL descriptor its rank order position in the hierarchy of ACTFL descriptors. Thus, the equal-interval scale value of the ACTFL ratings novice-low, novice-intermediate, novice-high, intermediate-low, etc. were assigned the numbers 1, 2, 3, 4, etc. The highest rating, distinguished, was assigned a 10 and was only used for listening and reading. (See Appendix A for the descriptors and the numbers assigned.)

Convenience was an obvious motive for using the equal-interval quantification of the ACTFL ratings; however, due to the arbitrary nature of the quantification, we also used the quantification of ACTFL levels that was developed by Lange and Lowe (1987). Figure 1 presents a plot of the Lange and Lowe scaling of the ACTFL levels. To facilitate comparison, the equal-interval scale is, in this figure only, adjusted so that its end-points and those of the Lange-Lowe scale coincide. Note that it regards the interval from novice-low to novice-mid (the bottom two levels, 0.1 to 0.5), as well as the interval from intermediate-low to intermediate-mid (the fourth and fifth levels, 1.3 to 1.7), as relatively short, with the larger differences occurring at the high end of the scale.

The off-diagonal entries of Tables 3a-g contain correlations of ACTFL ratings and TOEFL test scores for schools A-G. For each school's table, we averaged multiple ratings for each student. The first three rows and columns of Tables 3a-g refer to ratings; those above the leading diagonal are based on equal-interval scaling of ACTFL ratings and those below the diagonal are based on Lange-Lowe scaling. Rows and columns 4 through 7 index TOEFL scores. Thus, the entries in rows 1 through 3 and columns 4 through 7 (to the right of the first three diagonal entries) are correlations of TOEFL scores with ACTFL ratings quantified using the equal-interval scale; those in columns 1 through 3 and rows 4 through 7 (below the first three diagonals) are correlations of TOEFL scores with ACTFL ratings quantified using the Lange-Lowe scale. Thus, the value .77 in row 1 column 4 of Table 3a was the value of the correlation between ACTFL ratings, quantified on the equal-interval scale, and TOEFL listening. The symmetrically placed value of .81 in row 4 and column 1 of Table 3a was the value of the correlation between ACTFL, quantified on the Lange-Lowe scale, and TOEFL listening. Note that the two entries, the .77 and the .81, are both in boldface. These, like all of the boldfaced correlations in Tables 3a-g, give the value of correlations between tests and ratings on like skills--TOEFL listening with ACTFL listening, TOEFL writing with ACTFL writing, or TOEFL reading with ACTFL reading.

The diagonal entries of Tables 3a-g contain the numbers of cases on which the correlations are based. For example, the 31 in the upper left-hand corner of Table 3a indicates th^ ., for School A, ratings on ACTFL listening for 31 students were available. Note that the first three diagonal entries differ. These differences result from the fact that raters rated only those students they felt able to evaluate and rated only those characteristics for which they felt qualified. The number of TOEFL scores available is given in leading diagonal entries 4 through 7. Not all raters were able to rate all characteristics, but all TOEFL scores were available for all students rated. Thus, the correlations in the last four columns of the first three rows were all based on the same number of students, and that number is the number in the corresponding row diagonal entry. The correlations in the bottom right-hand four-by-four section of the tables are based on all the students whose data were used for any rating. When two ratings were compared, the numbers of cases were the numbers of students for whom at least two ACTFL proficiencies were rated. These numbers, given in the table notes d, e, and f, are much smaller than the numbers on the diagonals. Thus, in Table 3a, 31 students were rated on the ACTFL listening scale and 39 were rated on the writing scale; but table note d indicates that only 16 students were rated on both scales. Because the numbers given in these footnotes--for example, the 16 for table 3a--are small, the correlations are especially subject to error.

By scanning the symmetrical entries in the sections of Tables 3a-g, one can compare the correlations based on equal-interval scaling with those based on the Lange-Lowe scaling. They are not identical, but they are not very different. Finding the minor differences between the two scalings was not surprising because the differences between identically ordered and reasonably spaced correlations are generally similar, being well approximated by the rank-order correlation. This is not to say that transforming the scale is not useful, only that reasonable transformations are not expected to affect correlations greatly.

The entries in rows 1 through 3 and columns 4 through 7 of Tables 3a-g correlation matrices are all substantial, as are the corresponding symmetrically placed entries. However, nothing requires them to be so because a student's test-taking behavior leading to a test score is different from that student's behavior leading to the rating. The test items elicit very restricted samples of behavior, a logistic requirement of large-volume test administration. In contrast, the ratings are based on teachers' observations of classroom performance, a much less restricted behavior sample. In the absence of data, one m' ht take the extreme view that the test scores and ratings should be uncorrelated because the behaviors elicited are so different. Our concept of the two measures predicts, however, that correlations between the two types of measures should be positive and substantial. The fact that the correlations were indeed positive and substantial contributes to our confidence in the rating, the test, and our concept that they reflect similar proficiencies.

Discriminant Validity

Tables 3a-g contain correlations that bear on the "discriminant validity" of the measures. This term refers to expectations about the sizes of correlations between measures of the same construct as compared to the sizes of correlations between measures of different constructs. The latter are expected to be lower if one's notion of which measures gauge which constructs is accurate. One might expect listening, writing, and reading to be somewhat different skills. One might also expect that the listening rating and the listening test score would correlate more highly than might the listening rating and the writing test score, for example. To facilitate such comparisons, correlations in which the rating and test score evaluated like-named skills appear in boldface type in Tables 3a-g.

The general trend as regards discriminant validity in Tables 3a-g is probably best appreciated by examining Table 4, where data from all schools are combined. All entries in this table were computed assuming equal intervals for the ACTFL proficiency levels. In this table, as in the previous Tables 3a-g, the correlations of measures of like skills are in boldface. The boldfaced .57, .55, and .61 are not exceptionally large as compared with the other entries. Tables 3a-g and 4 do not provide strong support for discriminant validity of the ratings and test scores.

Reliability of the Ratings

The reliability of the ratings is of concern in this study because it limits the correlations that can be obtained. It was possible to estimate reliability using the data of students for whom multiple ratings on the ACTFL scale were obtained. For these students, we computed a sum of squares around the ratee mean. These sums of squares were added for all students for whom such multiple ratings were obtained. The degrees of freedom is the number of ratings on such people minus the number of people (i.e., the number of observations less a degree of freedom for each ratee mean). The sum of squares/d.f. estimates a variance of the error, V(e). Then, if V(t) is the test variance and r is the reliability, the formula

$$r = (V(t) - \underline{W}\ V(e)) / V(t),$$

was used. The operation in the numerator of the equation removes the error variance from the test variance, leaving the remainder as the true score variance. The reliability thus estimated is the ratio of the true score variance to the test variance. The symbol $\underline{W}$ is included in the equation to provide for averaging or not averaging the ratings of students who were rated more than once. If the ratings were not averaged, as in Table 5a, then the value of $\underline{W}$ is one. However, if the data are averaged, there is relatively less error, and $\underline{W}$ is the average over ratees of the reciprocal of the number of raters per ratee.

Table 5a contains the total group reliabilities; Table 5b contains the reliabilities based on averaged ratings on the equal-interval scale. The reliabilities are in the far right columns in the Tables 5, with statistics based on the equal-interval scale adjacent to those based on the Lange-Lowe scale, which are in parentheses. Corresponding reliabilities from the two tables using all the total samples are .53 and .61 for listening, .61 and .64 for writing, and .62 and .63 for reading, respectively, on the equal-interval scale. Though these coefficients for the same modality are not greatly different for the total sample, larger differences may be noted for the individual schools. Thus, the School B listening reliability of .57 from Table 5a rises to .66 in Table 5b because of the averaging process used in Table 5b. All coefficients in Table 5b are larger than the corresponding ones in Table 5a because of the averaging process. The coefficients from Table 5b were used to correct the Table 4 total group correlations between tests and ratings of like characteristics for unreliability in the ratings. This was done, using equal-interval scale figures, by dividing the correlations between like characteristics, which are .57, .55, and .61 for listening, writing, and reading, respectively, by the square root of their reliabilities (Gulliksen, 1987) of the ratings. The resulting corrected correlations are .73, .69, and .77. Though these corrected correlations are substantial, their maxima are in the low 90s, hence, the ratings and tests are not entirely parallel. The relationships are, however, clearly quite substantial.

## Cross-Tabulation of ACTFL Ratings and TOEFL Section Scores

Table 6 contains a cross-tabulation of ACTFL ratings and TOEFL section scores. The distributions presented relate scores on measures of like constructs. That is, the top section gives the ACTFL listening rating against the TOEFL listening section score. Data used were the total numbers of ratings regardless of rater, ratee, or the institute attended by the ratee. Read the table as follows: the second column from the left in the upper section of the table shows that 10 (4+4+2) out of 21, or 48 percent, of the ratings given to persons with scores below 40 were at or below an ACTFL level of intermediate-low (I-L). (The descriptor for the Intermediate-low listening rating is given in Appendix A.) Note that the main mass of the distributions shifts up as one moves from left to right, so that high ratings go with high scores. This distribution shift from lower left to upper right reflects the correlation of ACTFL and TOEFL listening. Tables giving the cumulative distributions, to which users might be more accustomed, are given in Appendix C. Providing tables like those in Appendix C was an important goal of this research.

## Discussion

As we mentioned at the outset of this project, an attempt was made to improve on the ACTFL descriptors because of possible ambiguities they contain. We subsequently decided to use the descriptors as they currently exist, but there were some problems. One of these was that the descriptors sometimes referred to certain situations demanding very different levels of language proficiency. For example, the complexity of "spontaneous face-to-face conversations" can differ greatly depending on who is involved. Another problem noted was that the behaviors referred to by descriptors are not always clearly specified, as, for example, in distinguishing between "comprehends" and "understands" in the reading and listening descriptors. Reading and listening are receptive skills, not overt behaviors, but the rater can only infer from overt behavior whether comprehension or understanding has taken place. We might even say that the rating level is not completely defined until the behavior referents of the terms used are defined. As a practical matter, however, teaching the kinds of situations to which the rating levels apply, and the behavioral referents of the terms used, might require extensive sessions. The necessary communication might not be feasible within constraints imposed by the written descriptions of rating levels such as we used. Also, training the raters might result in mismatched rater and user interpretations of the descriptors because the users (deans, for example) will not be trained.

One problem that can arise if the definitions are not anchored in overt behavior is that the raters may fall back on their general impression of the ratees' language proficiency. The rating system will then be essentially a single-factor assessment of general proficiency. A second problem is that a local interpretation of the descriptor language could arise. We have noted in Table 2 the differences in the alignment of test scores and ratings at different institutions. More concrete and less relativistic language in the ACTFL guidelines might have prevented these differences. Having said this, we also suggest that if any new system is developed, an important part of its evaluation will be to see how it links to the present system. Indeed, the procedure by which any new system is developed should probably be data-based.

This project used raters not specifically trained in the use of the ACTFL descriptors to obtain evaluations of ESL students. The raters' task was to match ratee behavior to the behavior descriptions in the ACTFL proficiency levels. But the raters' lack of specific training in the use of the descriptors, together with possible ambiguities in the descriptors, suggested that raters' idiosyncratic tendencies to give high or low marks might override the required matching of proficiency levels. The finding that adjustments for rater difficulty within institutions did not affect the correlations is very supportive of other uses of the ACTFL descriptors. The usefulness of the descriptors is diminished if the tendencies of the person assigning ratings has to be taken into account. Possibly, more sensitive studies than this one could detect some degree of difference between raters. But the existence of small differences would not negate the value of the rating system.

The many substantial correlations between test scores and ratings provided additional evidence that the ACTFL ratings and the TOEFL scores tap similar underlying skills. Because the raters were untrained in the use of the ACTFL scales and because the specific behaviors required by multiple-choice tests are quite limited in contrast with the natural language performance noted in ESL Institutes, some readers might expect that correlations between the two types of measures would be near zero. That was not the case. The construct validity of both types of measures is thus enhanced. It should be noted, however, that even when corrected for unreliability of the ratings, these correlations were not perfect. The constructs supported are those of correlated skills rather than one-to-one correspondence. For example, the ratings of listening seem to be more directly related to interactive situations than are the items in the Listening Comprehension section of TOEFL. Hence, the skills may not be identical, but it is reasonable to expect them to be correlated, as the results showed.

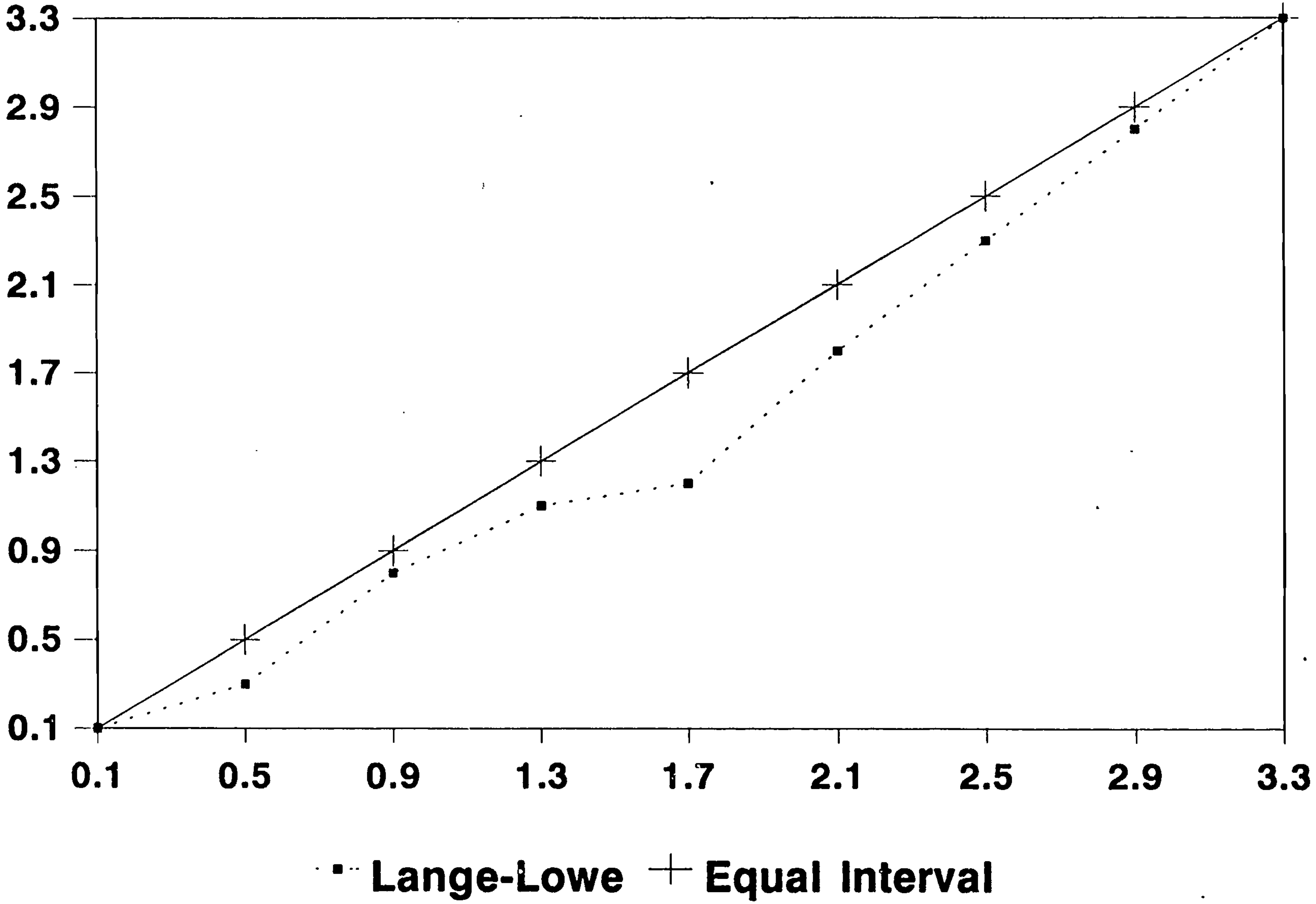
The evidence cited above supports the belief that TOEFL scores and ACTFL ratings reflect common skills. That is, the high correlations between reading, writing, and listening scores and the lack of evidence of discriminant validity might incline one to believe that only one construct underlying the rating-test domain can be supported. Indeed, some might insist that this is the case. But the high correlations between listening, writing, and reading skills could also result from correlated abilities to perform the skills, which are themselves different. Of course, assuming that different skills exist and result from correlated but different abilities is less parsimonious than assuming that listening, reading, and writing tasks all measure the same ability. Therefore, if we are to believe that these skills are different, we need to design situations in which their differences can be demonstrated. If many attempts to produce these situations fail, parsimony will eventually win out. One way of supporting the belief that the skills are different is to find individuals whose scores on the listening measures, for example, are high but whose scores on reading and writing are low, and for whom these differences are reliable. The individuals for whom such differences were noted could be reevaluated on parallel measures to test whether the differences recur or shrink away. If they recur, the notion that the constructs are different but correlated would be supported. Such examinees might be found among those scoring lower than most participants in this study, where differential skill development might be more common.

The data would be more supportive of our concepts of what the TOEFL section scores and the ratings measure if indications of discriminant validity were stronger. But aligning distributions of ACTFL descriptors with TOEFL score ranges can still serve a purpose. The distributions are informative in that they describe the relative frequency with which types of language-related performance occur at different TOEFL test score ranges. Though individual scores are subject to statistical variation, this information should be helpful to those responsible for interpreting TOEFL test scores.

Table 6 and the tables in Appendix C are important products of this research because, by using them and the ACTFL Guidelines, test users can bring more functional meaning to the TOEFL scores. In evaluating Table 6, it is important to note that the ratings and scores have substantial correlations--are even more substantial, though not perfect, if reliabilities are taken into account. Those who use the ratings should know that, within institutions, variations in raters' tendencies to be critical should not be a problem in using the table because the study has shown that rater severity or laxity does not affect the correlation with TOEFL section scores. However, the degree of relationship between scores and ratings in this table is somewhat attenuated by the institutional differences noted in Table 2. Institutions that choose to do so might sharpen the relationships between scores and ratings using locally collected data.

# Figure 1

## Lange-Lowe vs. Equal Interval Scales



## Table 1

Results of Correlating Ratings Adjusted for Rater Difficulty with Unadjusted Ratings and with TOEFL

### School C -- Listen

| r with TOEFL | Degrees Freedom | r with Rating | Type of Correction |
|---|---|---|---|
| 0.74 | 87 | 0.78 | Uncorrected |
| 0.76 | 79 | 0.92 | Football |
| 0.76 | 71 | 0.95 | Linear |

### School E -- Listen

| r with TOEFL | Degrees Freedom | r with Rating | Type of Correction |
|---|---|---|---|
| 0.71 | 154 | 0.88 | Uncorrected |
| 0.76 | 142 | 0.95 | Football |
| 0.74 | 130 | 0.96 | Linear |

### School E -- Write

| r with TOEFL | Degrees Freedom | r with Rating | Type of Correction |
|---|---|---|---|
| 0.74 | 63 | 0.86 | Uncorrected |
| 0.75 | 52 | 0.96 | Football |
| 0.74 | 41 | 0.97 | Linear |

### School E -- Read

| r with TOEFL | Degrees Freedom | r with Rating | Type of Correction |
|---|---|---|---|
| 0.60 | 34 | 0.97 | Uncorrected |
| 0.60 | 24 | 0.98 | Football |
| 0.60 | 14 | 0.99 | Linear |

## Table 2

Institutions' Means, Standard Deviations, and Numbers of Ratees and Raters for ACTFL Ratings[a] and TOEFL Scores

| | ACTFL Ratings | | | TOEFL Scores | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Institution A | | | | | | | |
| Mean | 6.16[b] | 6.59 | 6.97 | 49.06 | 46.99 | 47.80 | 479.48 |
| S.D. | 1.53 | 1.30 | 1.07 | 6.74 | 6.23 | 7.10 | 58.65 |
| N Ratees | 31 | 39 | 29 | 71 | 71 | 71 | 71 |
| N Raters | 6 | 5 | 5 | | | | |
| Institution B | | | | | | | |
| Mean | 5.97 | 6.69 | 6.80 | 48.04 | 48.13 | 47.73 | 479.69 |
| S.D. | 1.20 | 1.24 | 1.33 | 6.80 | 4.99 | 6.61 | 52.07 |
| N Ratees | 69 | 68 | 61 | 105 | 105 | 105 | 105 |
| N Raters | 12 | 8 | 10 | | | | |
| Institution C | | | | | | | |
| Mean | 5.58 | 5.55 | 5.14 | 49.15 | 47.15 | 46.18 | 474.85 |
| S.D. | 1.28 | 1.28 | 1.78 | 5.94 | 7.78 | 7.71 | 64.50 |
| N Ratees | 29 | 28 | 29 | 40 | 40 | 40 | 40 |
| N Raters | 8 | 4 | 4 | | | | |
| Institution D | | | | | | | |
| Mean | 6.42 | 6.20 | 6.51 | 51.43 | 49.75 | 49.88 | 503.53 |
| S.D. | 1.56 | 1.28 | 1.34 | 5.75 | 6.08 | 6.78 | 53.50 |
| N Ratees | 51 | 45 | 43 | 77 | 77 | 77 | 77 |
| N Raters | 10 | 9 | 9 | | | | |
| Institution E | | | | | | | |
| Mean | 6.06 | 6.14 | 6.13 | 53.68 | 48.57 | 48.25 | 501.68 |
| S.D. | 1.59 | 1.26 | 1.47 | 6.23 | 6.74 | 6.59 | 57.94 |
| N Ratees | 102 | 93 | 99 | 117 | 117 | 117 | 117 |
| N Raters | 12 | 11 | 10 | | | | |
| Institution F | | | | | | | |
| Mean | 5.21 | 6.54 | 6.27 | 49.69 | 46.80 | 46.88 | 477.86 |
| S.D. | 1.70 | 1.34 | 1 61 | 7.03 | 8.16 | 6.92 | 67.30 |
| N Ratees | 49 | 54 | 41 | 88 | 88 | 88 | 88 |
| N Raters | 6 | 6 | 4 | | | | |
| Institution G | | | | | | | |
| Mean | 7.57 | 6.61 | 6.96 | 47.35 | 45.01 | 44.72 | 456.91 |
| S.D. | 1.56 | 1.30 | 1.56 | 6.43 | 6.85 | 6.78 | 59.17 |
| N Ratees | 75 | 78 | 67 | 89 | 89 | 89 | 89 |
| N Raters | 10 | 10 | 9 | | | | |
| Total | | | | | | | |
| Mean | 6.13 | 6.38 | 6.43 | 49.95 | 47.55 | 47.43 | 483.12 |
| S.D. | 1.60 | 1.32 | 1.55 | 6.84 | 6.80 | 7.00 | 60.65 |
| N Ratees | 405 | 405 | 369 | 587 | 587 | 587 | 587 |
| N Raters | 64 | 53 | 51 | | | | |

[a] The ratee's average equal-interval scale rating was used.

[b] Levels are as follows: 5 is Intermediate-Mid, 6 is Intermediate-High, and 7 is Advanced.

ERIC
Full Text Provided by ERIC

### Table 3a

Correlations[a] among ratings and test scores
for Institution A

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 31[c] | .37[d] | .64[e] | **.77** | .66 | .73 | .78 |
| Write | .35 | 39 | .41[f] | .42 | **.33** | .27 | .41 |
| Read | .65 | .39 | 29 | .61 | .62 | **.63** | .69 |
| | | | | | | | |
| Listen | **.81** | .43 | .60 | 71 | .66 | .66 | .89 |
| Write | .67 | **.33** | .63 | | 71 | .63 | .86 |
| Read | .74 | .28 | **.62** | | | 71 | .88 |
| Total | .80 | .41 | .69 | | | | 71 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the equal-interval scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=16 [e] N=17 [f] N=17

### Table 3b

Correlations[a] among ratings and test scores
for Institution B

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 69[c] | .25[d] | .49[e] | **.60** | .31 | .39 | .50 |
| Write | .35 | 68 | .55[f] | .22 | **.48** | .51 | .52 |
| Read | .65 | .39 | 61 | .39 | .67 | **.79** | .74 |
| | | | | | | | |
| Listen | **.61** | .25 | .39 | 105 | .57 | .68 | .90 |
| Write | .35 | **.49** | .67 | | 105 | .45 | .76 |
| Read | .43 | .52 | **.79** | | | 105 | .86 |
| Total | .54 | .54 | .74 | | | | 105 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the equal-interval scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=53 [e] N=48 [f] N=49

## Table 3c

Correlations[a] among ratings and test scores
for Institution C

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 29[c] | .75[d] | .89[e] | **.76** | .63 | .58 | .70 |
| Write | .73 | 28 | .89[f] | .76 | **.83** | .70 | .83 |
| Read | .86 | .88 | 29 | .78 | .83 | **.74** | .86 |
| Listen | **.74** | .73 | .77 | 40 | .65 | .61 | .81 |
| Write | .61 | **.82** | .84 | | 40 | .86 | .94 |
| Read | .57 | .69 | **.74** | | | 40 | .93 |
| Total | .70 | .82 | .85 | | | | 40 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the ordinal scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=28 [e] N=28 [f] N=28

## Table 3d

Correlations[a] among ratings and test scores
for Institution D

| | | Rating[b] | | | Test | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 51[c] | .44[d] | .29[e] | **.61** | .37 | .44 | .55 |
| Write | .40[g] | 45 | .88[f] | .51 | **.50** | .51 | .59 |
| Read | .31[g] | .86[g] | 43[g] | .54 | .55 | **.48** | .61 |
| Listen | **.61** | .49 | .52 | 77 | .56 | .52 | .79 |
| Write | .39 | **.48** | .49 | | 77 | .75 | .90 |
| Read | .45 | .47 | **.41** | | | 77 | .89 |
| Total | .56 | .57 | .56 | | | | 77 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the ordinal scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=41 [e] N=39 [f] N=36

[g] Because the Lange-Lowe scale has no value for a rating of "Distinguished," one case was lost, leaving 42 for correlation with the tests. The N's corresponding to footnotes d,e, and f, but on the Lange-Lowe scale, are 41, 38, and 35.

ERIC
Full Text Provided by ERIC

## Table 3e

Correlations[a] among ratings and test scores for Institution E

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 102[c] | .69[d] | .80[e] | **.71** | .68 | .62 | .76 |
| Write | .68 | 93 | .73[f] | .71 | **.72** | .57 | .76 |
| Read | .80 | .75 | 99 | .62 | .65 | **.60** | .72 |
| Listen | **.69** | .71 | .62 | 117 | .69 | .68 | .89 |
| Write | .66 | **.74** | .66 | | 117 | .68 | .89 |
| Read | .62 | .59 | **.62** | | | 117 | .89 |
| Total | .75 | .77 | .72 | | | | 117 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the ordinal scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=89 [e] N=95 [f] N=89

## Table 3f

Correlations[a] among ratings and test scores for Institution F

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 49[c] | .67[d] | .64[e] | **.71** | .72 | .69 | .79 |
| Write | .66 | 54 | .58[f] | .66 | **.74** | .65 | .76 |
| Read | .63 | .53[g] | 41[g] | .68 | .74 | **.71** | .77 |
| Listen | **.72** | .66 | .67 | 88 | .77 | .78 | .92 |
| Write | .71 | **.74** | .70 | | 88 | .71 | .91 |
| Read | .71 | .67 | **.71** | | | 88 | .90 |
| Total | .79 | .76 | .76 | | | | 88 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the ordinal scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=44 [e] N=33 [f] N=38

[g] Because the Lange-Lowe scale has no value for a rating of "Distinguished," one case was lost leaving 42 for correlation with the tests. The N corresponding to footnote f was 37.

**Table 3g**

Correlations[a] among ratings and test scores for Institution G

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 75[c,g] | .71[d] | .88[e] | **.71** | .73 | .50 | .74 |
| Write | .68[g] | 78[g] | .82[f] | .57 | **.69** | .65 | .72 |
| Read | .83[g] | .80[g] | 67[g] | .74 | .79 | **.72** | .83 |
| | | | | | | | |
| Listen | **.69** | .57 | .71 | 89 | .72 | .60 | .87 |
| Write | .69 | **.72** | .74 | | 89 | .70 | .91 |
| Read | .41 | .66 | **.64** | | | 89 | .87 |
| Total | .69 | .73 | .78 | | | | 89 |

[a] Where data points for the correlations of ratings are based on multiple ratings, the ratings are averaged.

[b] Ratings for correlations above the leading diagonal are expressed on the ordinal scale; they are expressed on the Lange-Lowe scale for correlations below the leading diagonal.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=65 [e] N=54 [f] N=66

[g] Because the Lange-Lowe scale has no value for a rating of "Distinguished," three cases were lost for Listening, and three for Reading, leaving 72 and 64 respectively for correlation with the tests. The N's corresponding to footnotes d,e, and f, but on the Lange-Lowe scale, are 62, 51, and 63.

ERIC Full Text Provided by ERIC

## Table 4

Correlations[a] among ratings and test scores
Total Group

| | Rating[b] | | | Test | | | |
|---|---|---|---|---|---|---|---|
| | Listen | Write | Read | Listen | Write | Read | Total |
| Listen | 406[c] | .58[d] | .70[e] | **.57** | .61 | .52 | .65 |
| Write | .58 | 405 | .73[f] | .49 | **.55** | .52 | .59 |
| Read | .70 | .73 | 369 | .53 | .60 | **.61** | .66 |
| Listen | **.57** | .49 | .53 | 587 | .66 | .65 | .87 |
| Write | .61 | **.55** | .60 | .66 | 587 | .68 | .88 |
| Read | .52 | .52 | **.61** | .65 | .68 | 587 | .88 |
| Total | .65 | .59 | .66 | .87 | .88 | .88 | 587 |

[a] Where data points for the ratings' correlations are based on multiple ratings, the ratings are averaged.

[b] All ratings were on the ordinal scale.

[c] Underlined diagonal entries are the numbers of cases available for correlations with test scores.

[d] N=336 [e] N=314 [f] N=323

## Table 5a

Reliability Analysis for Ratings Not Averaged[a]

| | | --------Variances------- | | | Deg. Free. |
|---|---|---|---|---|---|
| ACTFL Scale | School | Rating | Error of Meas. | Reliab. | Err. Meas. |
| Listen | B | 1.45(.30)[b] | .62(.11) | .57(.63) | 30 |
| | C | 3.02(.50) | 1.63(.28) | .46(.44) | 92 |
| | E | 2.90(.53) | 1.21(.24) | .58(.55) | 155(152) |
| | Total | 2.86(.52) | 1.35(.25) | .53(.52) | 306(303) |
| Write | C | 2.01(.36) | .29(.06) | .86(.83) | 30 |
| | E | 1.59(.33) | .65(.11) | .59(.67) | 66 |
| | Total | 1.83(.39) | .72(.13) | .61(.67) | 116 |
| Read | C | 3.14(.49) | .78(.16) | .75(.67) | 37 |
| | E | 2.21(.42) | .51(.08) | .77(.81) | 34 |
| | Total | 2.61(.48) | .99(.20) | .62(.58) | 81 |

[a] Reliabilities computed with W equal to one.
[b] All figures in parentheses refer to the Lange-Lowe Scale.

## Table 5b

Reliability Analysis For Ratings Averaged[a]

| | | --------Variances------- | | |
|---|---|---|---|---|
| ACTFL Scale | School | Avg. Rating | Error of Meas.[c] | Reliab. |
| Listen | B | 1.44(.30)[b] | .62(.11) | .66(.71) |
| | C | 1.64(.27) | 1.63(.28) | .74(.73) |
| | E | 2.53(.46) | 1.21(.24) | .78(.76) |
| | Total | 2.56(.49) | 1.35(.25) | .61(.62) |
| Write | C | 1.64(.28) | .29(.06) | .95(.89) |
| | E | 1.59(.32) | .65(.11) | .72(.76) |
| | Total | 1.74(.37) | .72(.13) | .64(.69) |
| Read | C | 3.17(.46) | .78(.16) | .88(.83) |
| | E | 2.16(.42) | .51(.08) | .81(.84) |
| | Total | 2.40(.50) | .99(.20) | .63(.64) |

[a] Reliabilities computed with W as the average over ratees of the reciprocal of the number of raters per ratee.
[b] All figures in parentheses refer to the Lange-Lowe Scale.
[c] Degrees of freedom for the error of measurement are the same in Tables 5a and 5b.

## Table 6

### Frequency Distributions of ACTFL Rating for Several TOEFL Score Ranges

### ACTFL vs TOEFL -- Listening

| ACTFL | TOEFL Listening Score Range | | | | | | No. |
|---|---|---|---|---|---|---|---|
| Listen | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 | Ratings |
| Dist[3] | -- | -- | -- | -- | 3 | 3 | 6 |
| Sup | -- | -- | 6 | 12 | 15 | 9 | 42 |
| Adv+ | -- | 3 | 9 | 33 | 28 | 17 | 90 |
| Adv | -- | 10 | 43 | 58 | 32 | 13 | 156 |
| I-H | 3 | 27 | 44 | 43 | 34 | 3 | 154 |
| I-M | 8 | 41 | 50 | 34 | 12 | 2 | 147 |
| I-L | 2 | 33 | 29 | 10 | 5 | -- | 79 |
| N-H | 4 | 12 | 10 | 2 | 1 | -- | 29 |
| N-M | 4 | 11 | 5 | -- | 1 | -- | 21 |
| N-L | -- | 1 | -- | -- | -- | -- | 1 |
| Total | 21 | 138 | 196 | 192 | 131 | 47 | 725 |

### ACTFL vs TOEFL -- Writing

| ACTFL | TOEFL Writing Score Range | | | | | | No. |
|---|---|---|---|---|---|---|---|
| Write | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 | Ratings |
| Sup[3] | -- | -- | 1 | 4 | 8 | 7 | 20 |
| Adv+ | 1 | 6 | 20 | 26 | 30 | 5 | 88 |
| Adv | 3 | 23 | 46 | 35 | 24 | 1 | 132 |
| I-H | 13 | 28 | 53 | 40 | 5 | -- | 139 |
| I-M | 30 | 33 | 20 | 9 | 9 | -- | 101 |
| I-L | 14 | 17 | 7 | 2 | 1 | -- | 41 |
| N-H | 6 | 1 | -- | 1 | -- | -- | 8 |
| N-M | 1 | -- | -- | -- | -- | -- | 1 |
| N-L | -- | -- | -- | -- | -- | -- | -- |
| Total | 68 | 108 | 147 | 117 | 77 | 13 | 530 |

### ACTFL vs TOEFL -- Reading

| ACTFL | TOEFL Reading Score Range | | | | | | No. |
|---|---|---|---|---|---|---|---|
| Read | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 | Ratings |
| Dist[3] | -- | -- | -- | -- | 5 | 1 | 6 |
| Sup | -- | 3 | 4 | 11 | 7 | 5 | 30 |
| Adv+ | 1 | 9 | 11 | 20 | 21 | 4 | 66 |
| Adv | 6 | 17 | 44 | 38 | 13 | 1 | 119 |
| I-H | 16 | 32 | 31 | 25 | 5 | 1 | 110 |
| I-M | 17 | 26 | 17 | 7 | 1 | -- | 68 |
| I-L | 11 | 11 | 5 | 3 | -- | -- | 30 |
| N-H | 11 | 5 | 4 | 1 | -- | -- | 21 |
| N-M | 7 | -- | -- | -- | -- | -- | 7 |
| N-L | -- | -- | -- | -- | -- | -- | -- |
| Total | 69 | 103 | 116 | 105 | 52 | 12 | 457 |

[3] The following abbreviations are used for the ACTFL levels: N-L, N-M, and N-H for novice low, mid, and high, respectively; I-L, I-M, and I-H for intermediate low, mid, and high, respectively; Adv and Adv+ for advanced and advanced-plus, respectively; Sup for superior; and Dist for distinguished.

## Appendix A

## ACTFL Guidelines

Note: Equal-interval scale values are given in roman numerals at the start of the descriptor. For example, roman one (I) in the line beginning "Novice-Low I. Understanding..." means that the equal-interval listening scale value for Novice-Low was one. Those descriptors with similar shorthand names, such as Novice-Low were assigned the same scale values for all three scales--listening, writing, and reading. This assignment is not intended to imply an equating of any sort.

### Generic Descriptions - Listening

These guidelines assume that all listening tasks take place in an authentic environment at a normal rate of speech using standard or near-standard norms.

Novice-Low

I. Understanding is limited to occasional isolated words, such as cognates, borrowed words, and high-frequency social conventions. Essentially no ability to comprehend even short utterances.

Novice-Mid

II. Able to understand some short, learned utterances, particularly where context strongly supports understanding speech is clearly audible. Comprehends some words and phrases from simple questions, statements, high-frequency commands and courtesy formulae about topics that refer to basic personal information or the immediate physical setting. The listener requires long pauses for assimilation and periodically requests repetition and/or a slower rate of speech.

Novice-High

III. Able to understand short, learned utterances and some sentence-length utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends words and phrases from simple questions, statements, high-frequency commands and courtesy formulae. May require repetition, rephrasing and/or a slowed rate of speech for comprehension.

Generic Descriptions - Listening (Con't)

Intermediate-Low

IV. Able to understand sentence-length utterances which consist of recombinations of learned elements in a limited number of content areas, particularly if strongly supported by the situational context. Content refers to basic personal background and needs, social conventions and routine tasks, such as getting meals and receiving simple instructions and directions. Listening tasks pertain primarily to spontaneous face-to-face conversations. Understanding is often uneven; repetition and rewording may be necessary. Misunderstandings in both main ideas and details arise frequently.

Intermediate Mid

V. Able to understand sentence-length utterances which consist of recombinations of learned utterances on a variety of topics. Content continues to refer primarily to basic personal background and needs, social conventions and somewhat more complex tasks, such as lodging, transportation, and shopping. Additional content areas include some personal interests and activities, and a greater diversity of instructions and directions. Listening tasks not only pertain to spontaneous face-to-face conversations but also to short routine telephone conversations and some deliberate speech, such as simple announcements and reports over the media. Understanding continues to be uneven.

Intermediate High

VI. Able to sustain understanding over longer stretches of connected discourse on a number of topics pertaining to different times and places; however, understanding is inconsistent due to failure to grasp main ideas and/or details. Thus, while topics do not differ significantly from those of an Advanced level listener, comprehension is less in quantity and poorer in quality.

## Generic Descriptions - Listening (Con't)

Advanced

VII. Able to understand main ideas and most details of connected discourse on a variety of topics beyond the immediacy of the situation. Comprehension may be uneven due to a variety of linguistic and extralinguistic factors, among which topic familiarity is very prominent. These texts frequently involve description and narration in different time frames or aspects, such as present, nonpast, habitual, or imperfective. Texts may include interviews, short lectures on familiar topics, and news items and reports primarily dealing with factual information. Listener is aware of cohesive devices but may not be able to use them to follow the sequence of though in an oral text.

Advanced-Plus

VIII. Able to understand the main ideas of most speech in a standard dialect; however, the listener may not be able to sustain comprehension in extended discourse which is propositionally and linguistically complex. Listener shows an emerging awareness of culturally implied meanings beyond the surface meanings of the text but may fail to grasp sociocultural nuances of the message.

Superior

IX. Able to understand the main ideas of all speech in a standard dialect, including technical discussion in a field of specialization. Can follow the essentials of extended discourse which is propositionally and linguistically complex, as in academic/professional settings, in lectures, speeches, and reports. Listener shows some appreciation of aesthetic norms of target language, of idioms, colloquialisms, and register shifting. Able to make inferences within the cultural framework of the target language. Understanding is aided by an awareness of the underlying organizational structure of the oral text and includes sensitivity for its social and cultural references and its affective overtones. Rarely misunderstands but may not understand excessively rapid, highly colloquial speech, or speech that has strong cultural references.

ERIC
Full Text Provided by ERIC

Distinguished

X. Able to understand all forms and styles of speech pertinent to personal, social, and professional needs tailored to different audiences. Shows strong sensitivity to social and cultural references and aesthetic norms by processing language from within the cultural framework. Texts include theater plays, screen productions, editorials, symposia, academic debates, public policy statements, literary readings, and most jokes and puns. May have difficulty with some dialects and slang.

Generic Descriptions--Writing

Novice-Low

I. Able to form some letters in an alphabetic system. In languages whose writing systems use syllabaries or characters, writer is able to both copy and produce the basic strokes. Can produce romanization of isolated characters, where applicable.

Novice-Mid

II. Able to copy or transcribe familiar words or phrases and reproduce some from memory. No practical communicative writing skills.

Novice-High

III. Able to write simple fixed expressions and limited memorized material and some recombinations thereof. Can supply information on simple forms and documents. Can write names, numbers, dates, own nationality, and other simple autobiographical information as well as some short phrases and simple lists. Can write all the symbols in an alphabetic or syllabic system or 50-100 characters or compounds in a character writing system. Spelling and representation of symbols (letters, syllables, characters) may be partially correct.

Intermediate-Low

IV. Able to meet limited practical writing needs. Can write short messages, postcards, and take down simple notes, such as telephone messages. Can create statements or questions within the scope of limited language experience. Material produced consists of recombinations of learned vocabulary and structures into simple sentences on very familiar topics. Language is inadequate to express in writing anything but elementary needs. Frequent errors in grammar, vocabulary, punctuation, spelling and in formation of nonalphabetic symbols, but writing can be understood by natives used to the writing of nonnatives.

## Generic Descriptions--Writing (Con't)

Intermediate-Mid

V. Able to meet a number of practical writing needs. Can write short, simple letters. Content involves personal preferences, daily routine, everyday events, and other topics grounded in personal experience. Can express present time or at least one other time frame or aspect consistently, e.g., nonpast, habitual, imperfective. Evidence of control of the syntax of noncomplex sentences and basic inflectional morphology, such as declensions and conjugation. Writing tends to be a loose collection of sentences or sentence fragments on a given topic and provides little evidence of conscious organization. Can be understood by natives used to the writing of nonnatives.

Intermediate-High

VI. Able to meet most practical writing needs and limited social demands. Can take notes in some detail on familiar topics and respond in writing to personal questions. Can write simple letters, brief synopses and paraphrases, summaries of biographical data, work and school experience. In those languages relying primarily on content words and time expressions to express time, tense, or aspect, some precision is displayed; their tense and/or aspect is expressed through verbal inflection, forms are produced rather consistently, but not always accurately. An ability to describe and narrate in paragraphs is emerging. Rarely uses basic cohesive elements, such as pronominal substitutions or synonyms in written discourse. Writing, though faulty, is generally comprehensible to natives used to the writing of nonnatives.

## Generic Descriptions--Writing (Con't)

**Advanced**

VII. Able to write routine social correspondence and join sentences in simple discourse of at least several paragraphs in length on familiar topics. Can write simple social correspondence, take notes, write cohesive summaries and resumes, as well as narratives and descriptions of a factual nature. Has sufficient writing vocabulary to express self simply with some circumlocution. May still make errors in punctuation, spelling, or the formation of nonalphabetic symbols. Good control of the morphology and the most frequently used syntactic structures, e.g., common word order patterns, coordination, subordination, but makes frequent errors in producing complex sentences. Uses a limited number of cohesive devices, such as pronouns, accurately. Writing may resemble literal translations from the native language, but a sense of organization (rhetorical structure) is emerging. Writing is understandable to natives not used to the writing of nonnatives.

**Advanced-Plus**

VIII. Able to write about a variety of topics with significant precision and in detail. Can write most social and informal business correspondence. Can describe and narrate personal experiences fully but has difficulty supporting points of view in written discourse. Can write about the concrete aspects of topics relating to particular interests and special fields of competence. Often shows remarkable fluency and ease of expression, but under time constraints and pressure writing may be inaccurate. Generally strong in either grammar or vocabulary, but not in both. Weakness and unevenness in one of the foregoing or in spelling or character writing formation may result in occasional miscommunication. Some misuse of vocabulary may still be evident. Style may still be obviously foreign.

## Generic Descriptions--Writing (Con't)

Superior

IX. Able to express self effectively in most formal and informal writing on practical, social and professional topics. Can write most type of correspondence, such as memos as well as social and business letters, and short research papers and statements of position in areas of special interest or in special fields. Good control of a full range of structures, spelling, or nonalphabetic symbol production, and a wide general vocabulary allow the writer to hypothesize and present arguments or points of view accurately and effectively. An underlying organization, such as chronological ordering, logical ordering, cause and effect, comparison, and thematic development is strongly evident, although not thoroughly executed and/or not totally reflecting target language patterns. Although sensitive to differences in formal and informal style, still may not tailor writing precisely to a variety of purposes and/or readers. Errors in writing rarely disturb natives or cause miscommunication.

## Generic Descriptions-Reading

These guidelines assume all reading text to be authentic and legible.

Novice-Low

I. Able occasionally to identify isolated words and/or major phrases when strongly supported by context.

Novice-Mid

II. Able to recognize the symbols of an alphabetic and/or syllabic writing system and/or a limited number of characters in a system that uses characters. The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. Material understood rarely exceeds a single phrase at a time, and rereading may be required.

Novice-High

III. Has sufficient control of the writing system to interpret written language in areas of practical need. Where vocabulary has been learned, can read for instructional and directional purposes standardized messages, phrases or expressions, such as some items on menus, schedules, timetables, maps, and signs. At times, but not on a consistent basis,the Novice-High level reader may be able to derive meaning from material at a slightly higher level where context and/or extralinguistic background knowledge are supportive.

Intermediate-Low

IV. Able to understand main ideas and/or some facts from the simplest connected texts dealing with basic personal and social needs. Such texts are linguistically noncomplex and have a clear underlying internal structure, for example chronological sequencing. They impart basic information about which the reader has to make only minimal suppositions or to which the reader brings personal interest and/or knowledge. Examples include messages with social purposes or information for the widest possible audience, such as public announcements and short, straightforward instructions dealing with public life. Some misunderstandings will occur.

## Generic Descriptions-Reading (Con't)

Intermediate-Mid

V. Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs. Such texts are still linguistically noncomplex and have a clear underlying internal structure. They impart basic information about which the reader has to make minimal suppositions and to which the reader brings personal interest and/or knowledge. Examples may include short, straightforward descriptions of persons, places, and things written for a wide audience.

Intermediate-High

VI. Able to read consistently with full understanding simple connected texts dealing with basic personal and social needs about which the reader has personal interest and/or knowledge. Can get some main ideas and information from texts at the next higher level featuring description and narration. Structural complexity may interfere with comprehension; for example, basic grammatical reactions may be misinterpreted and temporal references may rely primarily on lexical items. Has some difficulty with the cohesive factors in discourse, such as matching pronouns with referents. While texts do not differ significantly from those at the Advanced level, comprehension is less consistent. May have to read material several times for understanding.

Advanced

VII. Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. The prose is predominantly in familiar sentence patterns. Reader gets the main ideas and facts and misses some details. Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language. Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader.

Generic Descriptions-Reading (Con't)

Advanced-Plus

VIII. Able to follow essential points of written discourse at the Superior level in areas of special interest or knowledge. Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or texts which treat unfamiliar topics and situations, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. An emerging awareness of the aesthetic properties of language and of its literary styles permits comprehension of a wider variety of texts, including literary. Misunderstandings may occur.

Superior

IX. Able to read with almost complete comprehension and at normal speed expository prose of unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on knowledge of the target culture. Reads easily for pleasure. Superior-level texts feature hypotheses, argumentation, and supported opinions and include grammatical patterns and vocabulary ordinarily encountered in academic/professional reading. At this level, due to the control of general vocabulary and structure, the reader is almost always able to match the meanings derived from knowledge of the language, allowing for smooth and efficient reading of diverse texts. Occasional misunderstanding may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. At the Superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text. (Top-down strategies rely on real-world knowledge and prediction based on genre and organizational scheme of the text. Bottom-up strategies rely on actual linguistic knowledge.) Material at this level will include a variety of literary texts, editorials, correspondence, general reports and technical material in professional fields. Rereading is rarely necessary, and misreading is rare.

ERIC
Full Text Provided by ERIC

Generic Descriptions-Reading (Con't)

Distinguished

X. Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all socio-linguistic and cultural references by processing language from within the cultural framework. Able to understand a writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader.

## Appendix B

Instructor's ID__________

# ACTFL

# PROFICIENCY GUIDELINES IN READING COMPREHENSION

## Instructor's Rating Booklet

These guidelines assume all reading texts to be authent:c and legible.

## INSTRUCTIONS

Using your class roster, select the student that you think is most proficient in reading comprehension. Read through the pages of descriptors until you find the one that most closely describes that student's level of proficiency. This need not be (and in fact, probably won't be) the highest level presented in this booklet. Clearly print the student's name on the first line below the descriptor that is the best fit.

Next, take the least proficient student in reading comprehension in your class and again reading through the set of descriptors, choose the one that most typifies that student's proficiency level. Clearly print that student's name on the first line.

The two selections made above form the extreme boundaries or anchors for your class. Now choose a typical middle-level student and find the descriptor that fits that student best. The descriptor should fall between the two anchors already defined by your first two choices.

Work through the remainder of your class roster alternating between high, low and middle-level students. For each student read the descriptors carefully to find one that best describes his/her proficiency in reading. It is important that students be placed according to the descriptors in the booklet regardless of class rank. Any descriptor may have appended to it as many students' names as you deem appropriate. Students will not necessarily be equally distributed across the categories.

EXAMPLE

| | |
|---|---|
| Advanced | VII. Able to read somewhat longer prose of several paragraphs...written for the general reader. [HERE APPEARED THE DESCRIPTOR FOR THE ADVANCED LEVEL IN READING--SEE APPENDIX A.] |

______________________________

______________________________

______________________________

EXAMPLE (Con't)

Novice-Mid

Able to recognize the symbols of an alphabetic and/or syllabic writing system and/or a limited number of characters in a system that uses characters. The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. Material understood rarely exceeds a single phrase at a time, and rereading may be required.

In the examples above, the student in class Z with the highest level of proficiency in reading was best described by ACTFL level "Advanced." Her name appears on the first line under that descriptor. The student ranked as the least proficient in reading was best represented by the description that appears in the booklet for "Novice-Mid." His name appears on the first line under that descriptor. The names of all the other students in class Z cannot fall outside of these two categories in the reading booklet. Please note, these are examples only. Your own extreme categories may be different.

If a student cannot be placed because you do not know the student well enough to make this rating, note this next to the student's name on the roster. Do not try to force a rating.

When you have rated all the students you can, review your placement for consistency. If you feel that changes are needed, simply cross out the misplaced name and enter it on the page you feel would be more appropriate.

Please return this booklet, along with the others, to the director of the intensive language program, along with a copy of your class roster. Thank you.

Novice-Low

Able occasionally to identify isolated words and/or major phrases when strongly supported by context.

Novice-Mid

Able to recognize the symbols of an alphabetic and/or syllabic writing system and/or a limited number of characters in a system that uses characters. The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. Material understood rarely exceeds a single phrase at a time, and rereading may be required.

## Appendix C

### Percentile Distributions of ACTFL Ratings for Several TOEFL Section Score Ranges

TOEFL scores would be more meaningful if they were aligned with verbal descriptors of language proficiency levels. The data below provide such an alignment. The descriptors of proficiency levels used in the ACTFL rating system were chosen for this purpose.

Examination of the ACTFL descriptors will reveal that they are broadly descriptive but not precise regarding any particular class of skills. The data given here are not a substitute for a careful assessment of specific skills where such are needed.

The tables are based on ratings of students' levels of language proficiency by instructors in ESL institutes and on TOEFL scores. The ratings were obtained using the ACTFL listening, writing, and speaking scales with directions like those given in Appendix B to this report. Training for the raters was limited to brief orientations using the written directions. Specific entries are interpreted in the table notes.

The tables contain two types of entries. First are cumulative percentiles of ratings for specified TOEFL ranges. The tabulations are conditioned on the TOEFL scores because the scores will be available to many TOEFL users before they observe examinees. They can use this information about TOEFL scores to make inferences about probable ACTFL proficiency levels, but not the other way around.

The percentiles are cumulated from low ratings to high ratings. That is, the second line of entries in each table gives the percent of those in the score range indicated at the top of the column as "Superior" or below ("Superior" is the highest). These entries appear in the upper portions of the tables.

The bottom lines of the tables give the numbers of ratees on which the columns' percentiles are based. The frequencies for any given column in the tables are not large; the frequencies for the extreme TOEFL intervals are quite small. Hence, the specifics of the distributions by rating categories are not well determined. However, the trends in the tables are clear and consistent. Therefore, using Table C-1, it is reasonable to conclude that an examinee whose Listening Comprehension score is below 40 on Listening is probably in the lower portion of those rated Intermediate on the ACTFL listening scale; one scoring above 59 would probably be rated Advanced Plus or better.

Obviously, the tables in this appendix apply only to TOEFL-takers. But it should be pointed out that there was some

systematic exclusion of students from the testing. Students perceived by the schools to be "unready for the TOEFL" were not required to take it. This decision was not made with ACTFL anchors in hand, but we suggest that those omitted came from the lower portion of the tables. If so, differences in the effective selectivity of schools in requiring the test could produce vertical shifts in the frequency distributions relative to the test, perhaps like those noted in Table 2 of this report.
For this reason, institutions might find it useful to rate the proficiency of examinees just before they take the TOEFL and compare their ratings with the scores. As these local data are accumulated, they will support more accurate interpretations of the tables.

## Table C-1

Percentile Distributions of ACTFL Ratings 4/
for Several TOEFL Listening Comprehension Score Ranges

| ACTFL | TOEFL Listening Score Range | | | | | |
|---|---|---|---|---|---|---|
| Listen | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 |
| Dist | 100 | 100 | 100 | 100 | 100 | 100 |
| Sup | 100 | 100 | 100 | 100 | 98 | 94 |
| Adv+ | 100 | 100 | 97 | 94 | 86 | 74 |
| Adv | 100 | 98 | 92 | 77 | 65 | 38 |
| I-H | 100 | 91 | 70 | 46 | 40 | 11 |
| I-M | 86 | 71 | 48 | 24 | 15 | 4 |
| I-L | 48[a] | 41 | 22 | 6 | 5 | -- |
| N-H | 38 | 17 | 8 | 1 | 2 | -- |
| N-M | 19 | 9 | 3 | -- | 1 | -- |
| N-L | -- | 1 | -- | -- | -- | -- |
| No. | 21 | 138[b] | 196 | 192 | 131 | 47 |

[a] The entry indicates that 48% of those who scored below 40 in Listening Comprehension received ACTFL listening ratings no higher than intermediate low.

[b] The entry indicates that 138 ratees achieved a TOEFL Listening Comprehension score from 40 to 44, inclusive. The total rated on this section of TOEFL was 725.

4/The following abbreviations are used for the ACTFL levels: N-L, N-M, and N-H for novice low, mid, and high, respectively; I-L, I-M, and I-H for intermediate low, mid, and high, respectively; Adv and Adv+ for advanced and advanced-plus, respectively; Sup for superior; and Dist for distinguished.

ERIC
Full Text Provided by ERIC

## Table C-2

Cumulative Frequency Distributions of ACTFL Ratings 5/
for Several TOEFL Structure and Written
Expression Score Ranges

| ACTFL | TOEFL Writing Score Range | | | | | |
|---|---|---|---|---|---|---|
| Write | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 |
| Sup | 100 | 100 | 100 | 100 | 100 | 100 |
| Adv+ | 100 | 100 | 99 | 97 | 90 | 46 |
| Adv | 99 | 94 | 86 | 74 | 51 | 8 |
| I-H | 94 | 73 | 54 | 44 | 19 | -- |
| I-M | 75 | 47 | 18 | 10 | 13[a] | -- |
| I-L | 31 | 17 | 5 | 3 | 1 | -- |
| N-H | 10 | 1 | -- | 1 | -- | -- |
| N-I | 1 | -- | -- | -- | -- | -- |
| N-L | -- | -- | -- | -- | -- | -- |
| No. | 68[b] | 108 | 147 | 117 | 77 | 13 |

[a] The entry indicates that 13% of those who scored from 55 to 59, inclusive, on TOEFL Structure and Written Expression received ACTFL writing ratings no higher than intermediate mid.

[b] The entry indicates that 68 ratees achieved a TOEFL Structure and Written Expression score below 40. The total number rated on this section of TOEFL was 530.

5/The following abbreviations are used for the ACTFL levels: N-L, N-M, and N-H for novice low, mid, and high, respectively; I-L, I-M, and I-H for intermediate low, mid, and high, respectively; Adv and Adv+ for advanced and advanced-plus, respectively; Sup for superior; and Dist for distinguished.

## Table C-3

Cumulative Frequency Distributions of ACTFL Ratings 6/
for Several TOEFL Vocabulary and Reading
Comprehension Score Ranges

| ACTFL | TOEFL Reading Score Range | | | | | |
|---|---|---|---|---|---|---|
| Read | <40 | 40-44 | 45-49 | 50-54 | 55-59 | >59 |
| Dist | 100 | 100 | 100 | 100 | 100 | 100 |
| Sup | 100 | 100 | 100 | 100 | 90 | 92 |
| Adv+ | 100 | 97 | 97 | 90 | 77 | 50 |
| Adv | 99 | 88 | 87 | 70 | 37 | 17 |
| I-H | 90 | 72 | 49 | 34 | 12 | 8 |
| I-M | 67 | 41 | 22 | 10 | 2 | -- |
| I-L | 42 | 16 | 8 | 4 | --[a] | -- |
| N-H | 26 | 5 | 3 | 1 | -- | -- |
| N-I | 10 | -- | -- | -- | -- | -- |
| N-L | -- | -- | -- | -- | -- | -- |
| No. | 69 | 103 | 116 | 105 | 52 | 12[b] |

[a] The dashes indicate that none of those who scored from 55 to 59, inclusive on TOEFL Vocabulary and Reading Comprehension received ACTFL reading ratings at intermediate mid or below.

[b] The entry indicates that 12 ratees achieved a TOEFL Structure and Written Expression score above 59. The total number rated on this section of TOEFL was 457.

6/The following abbreviations are used for the ACTFL levels: N-L, N-M, and N-H for novice low, mid, and high, respectively; I-L, I-M, and I-H for intermediate low, mid, and high, respectively; Adv and Adv+ for advanced and advanced-plus, respectively; Sup for superior; and Dist for distinguished.

# References

ACTFL (1986). ACTFL Proficiency Guidelines. Hastings-on-Hudson, NY: Author

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. Studies in Second Language Acquisition, 2, 149-164.

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. The Modern Language Journal, 70, 380-390.

Dandanoli, P. & Henning, G. H. (1990). An investigation of the construct validity of the ACTFL Proficiency Guidelines and Oral Interview Procedure. Foreign Language Annals, 12, No. 1, 11-21.

Douglas, Dan (1988). Testing listening comprehension in the context of the ACTFL Proficiency Guidelines. Studies in Second Language Acquisition, 2, 245-261.

ETS (1982). ETS Oral Proficiency Testing Manual. Princeton, NJ: Author.

Gulliksen, H. (1987) Theory of mental tests. Hillsdale, NJ: Lawrence-Erlbaum Associates.

Interagency Language Roundtable. (July, 1985). Language skill level descriptions. (Available from ACTFL Materials Center, Hustings-on-Hudson, NY).

Lange, D. L. & Lowe, Jr., P. L. (1987). Grading reading passages according to the ACTFL/ETS/ILR Reading Proficiency Standard: Can it be learned? Selected Papers from the 1986 Language Testing Research Colloquium. Monterey, California: Defense Language Institute, pp. 111-127.

Lantolf, J. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. The Modern Language Journal, 69, 337-345.

Lantolf, J. & Frawley, W. (1988). Proficiency: Understanding the construct. Studies in Second Language Acquisition, 2, 181-196.

Lowe Jr., Pardee (1988). The Unassimilated History. In Lowe, Jr., P. and Stansfield, C. W. (Eds.), Second language proficiency assessment: Current issues. Englewood Cliffs, NJ: Prentis Hall.

ERIC Full Text Provided by ERIC

## References (con't)

Savignon, S. J. (1985). Evaluation of communicative competence: The ACTFL provisional proficiency guidelines. *Modern Language Journal*, *69*, 129-134.

ERIC Full Text Provided by ERIC

BEST COPY AVAILABLE

57906-01205 • Y112M.5 • 275629 • Printed in U.S.A.