

**RESEARCH
REPORT**

September 1999
ETS RR-99-23

**Validating a Test
Designed to Assess ESL Proficiency
at Lower Developmental Levels**

Kenneth M. Wilson



Statistics & Research Division
Princeton, NJ 08541

ABSTRACT

This exploratory study examined relationships between a test designed to assess English-language listening comprehension and reading skills in samples of nonnative-English speakers at lower levels of developed proficiency in English as a second or foreign language (ESL or EFL). The test--called ESL-EZY--was developed by using items similar to but easier, on the average, than those being used in an existing ESL proficiency test designed for intermediate- or higher-level ESL users/learners. This paper reports evidence regarding the relationship between ESL-EZY scores and teachers' ratings of oral English proficiency in samples--assessed in diverse settings in the United States and Japan--selected to include subgroups that tend to differ relatively widely in average level of developed English proficiency. Scores on ESL-EZY were found to correlate relatively strongly with teachers' ratings within the respective subgroups; correlations were especially strong in relatively less proficient subgroups, suggesting that for younger ESL-students and other ESL-learners with limited developed functional proficiency in English as a foreign or second language, a test embodying properties similar to those represented by ESL-EZY might provide useful supplementary assessment information.

Key words: ESL proficiency, teachers' ratings, lower-level testing

ACKNOWLEDGMENTS

This research could not have been completed without the generous cooperation of teachers of English as a second language—in diverse, principally secondary school, settings in the United States, and in several English-language institutes in Japan—who administered the experimental ESL-proficiency measure employed in the study and provided ratings of the oral English proficiency of their students according to a behaviorally anchored rating schedule.

The rating schedule was prepared by Steven A. Stupak, and Matthew Sindlinger supervised the preparation and distribution of study materials to assessment sites which were identified primarily by Victor Thiry, director of the International Education Foundation in the United States, and Akira Ito, director of the Institute for International Business Communication in Japan. Katherine Pashley helped to develop the data file used in the study; Jay Breyer conducted analyses involved in equating two forms of the experimental test used in the study, and in transforming raw scores to a standard scale. Robert Boldt, Judith Gelb, Gordon Hale, and Don Powers provided helpful reviews of drafts of the study report. The Research Division at Educational Testing Service provided essential indirect support.

BACKGROUND

The TOEIC (Test of English for International Communication) test, which measures English language listening comprehension and reading comprehension skills, was developed by Educational Testing Service (ETS) at the invitation of the Japanese Ministry of International Trade and Industry (MITI), and introduced in Japan in 1979. TOEIC program activities currently are administered under the aegis of the Chauncey Group International, Princeton, NJ (USA). The test is used primarily, though not exclusively, in corporate settings throughout the world to assess English proficiency skills of nonnative-English speaking employees or prospective employees, most of whom are educated beyond the secondary-school level.

The TOEIC test is designed to discriminate effectively over a relatively wide range of developed ESL proficiency; TOEIC items, of course, are designed to be of "average difficulty" for the general test-taking population. The purpose of the present study was to assess the validity of a specially developed "easier" test—called ESL-EZY--made up of items similar to those in the TOEIC test, but less difficult, on the average. As will be described in detail later, evidence bearing on the validity of ESL-EZY scores was obtained by assessing their concurrent relationships with teachers' ratings of oral English proficiency rendered according to a behaviorally anchored schedule, in diverse settings in the United States and Japan, respectively. Samples assessed were composed primarily of secondary-level students--including subgroups at relatively low levels of developed proficiency in English as a second or foreign language, for which a test such as ESL-EZY might be expected to be a useful assessment instrument.

Given the procedures involved in its development, above, ESL-EZY logically should be expected to exhibit general validity-related properties similar to those that have been found to obtain for the TOEIC test itself. Accordingly, evidence bearing on the TOEIC test's validity is reviewed briefly, below, before turning to a detailed description of the present study.

Evidence Regarding TOEIC's Validity

In the initial TOEIC validation study, Woodford (1982) found high correlations (centering around $r = .80$) between scores on the TOEIC test and direct measures of reading, writing, and speaking ability, including the Language Proficiency Interview (LPI) procedure. After considering such evidence, the TOEIC test's content and other psychometric properties of the test, Perkins (1987, p. 82), offered the following summary conclusions in an independent technical review:

"In sum, TOEIC is a standardized, highly reliable and valid measure of English, specifically designed to assess real-life reading and listening skills of candidates who will use English in a work context. Empirical studies indicate that it is also a valid indirect measure of speaking and writing. The items assess major grammatical structures and reading skills and, in addition to being an integrative test, TOEIC also appears to tap communicative competence in that the items require the examinee to utilize his or her sociolinguistic and strategic competence" (p. 82).

Subsequent validation research provides further evidence of the TOEIC test's strong concurrent relationship with ratings of oral English proficiency (LPI ratings), in samples from

representative administrations in Japan (e. g., Wilson, 1989; Saegusa, 1985) and from comparable administrations in several other countries (Wilson, 1989). The latter study was undertaken specifically to establish and evaluate guidelines for inferring probable level of oral English proficiency from levels of performance on the TOEIC test, using a regression model.

Correlations between the TOEIC test and the LPI were consistently strong (centering around .7) in samples from France, Mexico, and Saudi Arabia, as well as Japan. There was substantial "fit" (indexed by the usual analysis of residuals) between average observed levels of LPI performance and levels estimated from scores on the TOEIC test across as well as within these samples when estimated levels were based on a general equation reflecting the regression of LPI ratings on scores on the TOEIC test in the combined sample. The foregoing pattern was also observed in a subsequent (internal) analysis of data for a sample from a TOEIC test administration in Thailand that arrived too late to be included in the study. ^a (see correspondingly lettered endnote)

Based on these findings, guidelines were developed for making actuarial inferences from scores on the TOEIC test regarding probable levels of LPI-assessed oral English proficiency in representative testing contexts. These guidelines indicate the level and range of oral English proficiency, behaviorally defined, that examinees within certain TOEIC-score ranges can be expected to exhibit.

The development and use of such guidelines, of course, does not indicate that the indirect and direct measures are "construct equivalent." However, statistically validated inferences about expected levels of performance on the direct measure can be drawn from knowledge of performance on the indirect measure, and vice versa.

For example, research involving native English-speaking students in the United States (e.g., Breland, 1977) indicates that ratings of writing samples obtained toward the end of the first year of college can be predicted as validly by preadmission scores on a multiple-choice test designed to measure knowledge of grammatical rules and conventions, as by ratings of actual preadmission writing samples. Such evidence suggests that screening for "writing ability" can be accomplished as effectively and validly, for screening purposes, by the cost-effective, multiple-choice test as by the relatively more expensive writing sample--not that the two approaches are equivalent in any structural sense.

Similarly, the TOEIC test's interpretive guidelines provide information that is useful in screening for LPI-assessed oral English proficiency.

THE PRESENT STUDY

The evidence reviewed above attests directly to the validity of the TOEIC Test. ESL-EZY's item-type kinship to the TOEIC test suggests that ESL-EZY is likely to exhibit similar properties. However, independent empirical evidence clearly is needed to assess the strength and consistency of relationships between ESL-EZY scores and criteria reflecting functional ability to use English, in samples from the targeted population of ESL users/learners.

The target population for a test such as ESL-EZY (or some comparable test) would include a relatively high proportion of ESL/EFL learners/users likely to be ratable at or below Level 1 on the ILR oral language proficiency scale. In such a population, ratings of proficiency according to the basic ILR scale would classify perhaps a majority of the examinees in only two or three proficiency categories. Accordingly, for purposes of validating ESL-EZY, it was important to obtain criterion ratings based on a scale permitting finer differentiation at lower levels of proficiency than is provided by the basic ILR scale. Such discrimination is provided in a modification of the ILR scale, developed by the American Council of Teachers of Foreign Languages (ACTFL) in collaboration with ETS (ETS, 1982), that has come to be known as the ACTFL Oral Proficiency Interview scale. Appendix A provides a brief overview of the rationale for this action, the full generic behavioral-level descriptions in the ACTFL modification, and an outline indicating the essential similarity as well as lower-level differences between the ACTFL and ILR scales.

The present study was undertaken to assess the strength and consistency of relationships between scores on ESL-EZY and ratings of "oral English proficiency" rendered by native English-speaking ESL teachers (in U.S. high-school settings) and native Japanese-speaking English teachers (in Japanese ESL-training contexts), according to levels described in a behaviorally anchored scale, with level descriptions drawn as direct or paraphrased excerpts from the corresponding ACTFL generic descriptions.

The rating schedule used in the study is shown in Table 1. It has the same number of levels as the generic ACTFL scale for speaking proficiency (shown in Appendix A, Exhibit A.1), but the level descriptions in the schedule generally do not reflect the full range of descriptive detail that is included in the corresponding generic descriptions (cf., Table 1 and Appendix A, Exhibit A.1). Ratings based solely on consideration of the behavioral descriptions in Table 1 are not necessarily equivalent to ratings based on the corresponding full generic descriptions, of course. This is an interesting question, but it is not at issue in the present study.

At the same time, the criterion measure of oral English proficiency employed in the present study differs from that employed in TOEIC-validation research reviewed briefly above, in ways which need to be considered in evaluating study findings. More specifically, in studies cited above, the criterion was ILR-scale-rated "oral English proficiency" as reflected in behavior elicited in formal Language Proficiency Interviews conducted by experienced, professional ESL interviewers/raters. In contrast, in this exploratory study, ratings were rendered by teachers of English as a second language, using a rating schedule with which they did not have previous experience to provide ratings based on "naturalistic observation" of the targeted linguistic behavior ("speaking proficiency" in regular ESL classroom settings); also both native English-speaking ESL teachers and nonnative English-speaking ESL teachers were involved in the ratings.

The present study was not designed to assess effects on validity possibly associated with the use of "informal" vs "formal" methods in obtaining ratings, nor was it designed to assess the comparability of ESL proficiency ratings for native and nonnative English-speakers. However, judging from the findings reported below, these issues warrant further inquiry.

Table 1

Teacher Assessment of Student's Ability

The descriptions below are in ascending order, with slight changes at each level as you read down the list. Check the box that represents the highest consistent or sustained level of ability for the student. You may want to consult with the student before preparing your reply.

1. STUDENT HAS NO ABILITY WHATSOEVER IN THE LANGUAGE.

2. STUDENT'S ORAL PRODUCTION IS LIMITED TO OCCASIONAL ISOLATED WORDS.

Student is unable to function in the language and essentially has no ability to communicate.

3. STUDENT'S VOCABULARY IS LIMITED TO THAT NECESSARY TO EXPRESS SIMPLE ELEMENTARY NEEDS AND BASIC COURTESY FORMULAE.

Student is able to operate only in a very limited capacity within very predictable areas of need. Syntax is fragmented and the majority of utterances consist of isolated words or memorized clusters. Utterances do not show evidence of creating with language or being able to cope with the simplest situations.

4. STUDENT IS ABLE TO SATISFY IMMEDIATE NEEDS.

Student has no real autonomy of expression, although there are some emerging signs of spontaneity and flexibility. Most utterances are telegraphic and word endings are often omitted, confused, or distorted. Student can ask questions or make statements with reasonable accuracy only when this involves short, memorized utterances or formulae.

5. STUDENT IS ABLE TO SATISFY BASIC SURVIVAL NEEDS AND MINIMUM COURTESY REQUIREMENTS.

In areas of immediate need or on very familiar topics, the student can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversation. Almost all utterances contain fractured syntax and other grammatical errors. There is little precision in information conveyed, due to the tentative state of grammatical development and little or no use of modifiers.

6. STUDENT DISPLAYS GRAMMATICAL ACCURACY IN BASIC CONSTRUCTIONS, E.G., SUBJECT-VERB AGREEMENT, PLURALS, BASIC WORD ENDINGS.

Student's vocabulary permits discussion of topics beyond basic survival needs, e.g., personal history, leisure-time activities. Student has an understanding of verb forms, but is not able to produce correct forms in conversation.

Table 1, Study Scale, continued

7. STUDENT HAS DEVELOPED FLEXIBILITY IN A RANGE OF CIRCUMSTANCES BEYOND IMMEDIATE SURVIVAL NEEDS. STUDENT SHOWS SPONTANEITY IN LANGUAGE PRODUCTION BUT FLUENCY IS VERY UNEVEN.

Student is able to satisfy linguistically most survival needs and limited social demands. Student can initiate and sustain a general conversation, but has little understanding of the social conventions of conversation. While some word order is established, errors still occur in more complex patterns. Longer utterances and unfamiliar subject matter are very difficult for the student to handle.

8. STUDENT'S ABILITY TO DESCRIBE AND GIVE PRECISE INFORMATION IS LIMITED. STUDENT IS AWARE OF BASIC COHESIVE FEATURES OF LANGUAGE, E.G., PRONOUNS, VERB INFLECTIONS.

Student's extended discourse is mainly composed of a series of short, discrete utterances. Student is able to satisfy routine social demands and limited work/study requirements. Student can handle with confidence, but not with facility, most social situations. This includes introductions and casual conversations about current events, as well as work, family and autobiographical information. Student has an active vocabulary sufficient to respond simply with some circumlocutions. Student can usually handle basic, high-frequency constructions quite accurately, but does not have thorough or confident control of higher level grammatical structures.

9. STUDENT EXHIBITS A STRONG ABILITY TO COMMUNICATE ON CONCRETE TOPICS RELATING TO PARTICULAR INTERESTS AND AREAS OF EXPERTISE.

10. STUDENT'S LEVEL IS ABOVE ALL DESCRIPTIONS PROVIDED HERE.

Student is able to satisfy most work/study requirements, even when certain complications are involved. Student often shows remarkable fluency and ease of speech, but under tension or pressure breakdown may occur. Areas of weakness in student's speech may range from occasional errors in high frequency constructions to repeated errors in more complex structures. Student has control of general vocabulary, although some groping for everyday vocabulary may still be evident.

Study Data and Procedures

ESL-EZY scores and ESL teachers' ratings were obtained for ESL students in scattered secondary school settings and one post-secondary school setting in the United States and in intensive programs that develop oral English proficiency in Japan. Students completed background questions, including questions on age and national origin.

The Sample

Subjects were classified, for study purposes, according to membership in four groups, described below. It was assumed a priori that Group 1 and Group 2 would function at higher average general levels of developed English proficiency than would Group 3 and Group 4.

Group 1 was composed of ESL users/learners, largely international students planning to study in a U.S. university, enrolled in an intensive ESL program at the University of Delaware, and tested and rated during the summer of 1987. They were nationally and linguistically heterogeneous, and their modal reported age-category 20 to 24 years.

Group 2 included only ESL secondary-level exchange students, studying under the auspices of the International Education Forum (IEF), tested and rated toward the beginning of their second semester in scattered secondary school settings in the United States, primarily during February and March, 1988. Most members of this nationally and linguistically heterogeneous sample were between 15 and 19 years of age and had been carefully screened for English proficiency as part of the IEF program's selection process. Group 3 was made up of Japanese secondary-level students between 15 and 19 years of age, in summer programs for the intensive study of English as a second language, tested and rated at three sites in the United States in late April/early May of 1988.

Group 4 involved Japanese examinees at diverse educational levels, who were enrolled in intensive ESL programs at several ECC (English Conversation Course) English-language institutes in Japan, and tested and rated by their native-Japanese speaking English language instructors primarily in July or September, 1988; their modal age-category was 20 to 24 years.

Study Data

Ratings for the respective groups were not obtained under strictly uniform conditions, due to the scattered sites and local control that characterized the testing and rating process. However, two general conditions of data collection were assumed to be common across all sites: local test administrations con-formed to the standard directions provided, and students were rated according to proficiency level after consideration by teacher/raters of the behavioral descriptions and brief instructions provided in the rating schedule (Exhibit A).^b

None of the ESL teachers involved had prior experience with the particular rating schedule employed. Based on information provided by the TOEIC Steering Committee in Japan, the ECC Institute students in Group 4 were in programs emphasizing oral English proficiency. They were rated by native Japanese-speaking ESL teachers who were probably reasonably well-acquainted with the students. The teachers did not base their ratings of oral English proficiency on behavior observed in isolated "interviews." This study does not assume that students in any of the groups involved were specially interviewed.

By inference, ESL teacher/raters for Groups 1, 2, and 4 had greater opportunity than did teacher/raters for Group 3 (summer program) to form impressions of oral English proficiency based on naturalistic observation. For example, Group 1 students when rated had been enrolled for some time in an ESL program at the University of Delaware; Group 2 (IEF Exchange) students were rated after a semester of study in their respective high schools, by ESL teachers who were familiar with them; and Group 4 students had been enrolled for some time in intensive ESL programs when rated by native Japanese-speaking ESL teachers. Group 3 students, on the other hand, were rated shortly after their arrival in the United States. The raters had only limited opportunity to observe

language proficiency during training, prior to rating the students. Based on the foregoing, rating conditions appear to have been somewhat less favorable in the summer program than in the other contexts involved in the study.

The ratings plausibly may tend to reflect cumulative impressions of oral English proficiency as observed by ESL teachers in daily interaction with students over differing periods of time, and in diverse contexts, in and out of the classroom. It is thus possible that the ESL teachers' ratings may reflect holistic impressions of developed levels of "general English proficiency"—that is, teachers' impressions may have been shaped not only by observation of behavior reflecting the ability to comprehend and produce ESL utterances (the exclusive focus in Language Proficiency Interviews), but also by observations of other aspects of proficiency (as exhibited in reading or writing samples, for example).

Multiple ratings were not obtained, thus precluding direct assessment of the reliability of the teachers' ratings, a consideration that was not at issue in the study.^c However, because the maximum correlation between two variables is limited by their respective reliabilities, relatively high correlations, if present, between teachers' ratings and ESL-EZY, provide pertinent indirect evidence bearing on the "pragmatic" reliability of teachers' ratings according to the schedule employed in the study.

Data Analysis

Raw scores for ESL-EZY sections from two experimental forms of the test were available, as were ratings according to the schedule shown in Table 1, for a total of 614 students. Using equating results provided by Breyer (J. Breyer, personal communications, Oct. 19, 1989; Nov. 1, 1989), ESL-EZY listening comprehension (LC-EZY), reading comprehension (R-EZY), and total (ESL-EZY) standard-scale scores were computed: LC-EZY and R-EZY were scaled in 5-point intervals (5, 10, 15, . . . , 90); ESL-EZY Total = LC-EZY + R-EZY (10, 15, 20, . . . , 180). See Appendix B for a brief description of the types of items used in the experimental test.

To assess the strength and consistency of ESL-EZY/Rating relationships, the regression of teachers' ratings on ESL-EZY scores was analyzed (a) in the total rated sample, (b) within each of the four groups described above, (c) in samples of Japanese and non-Japanese students, respectively, without regard to group, and (d) in subsamples classified by age-level across all groups. The regressions were also analyzed within Group 4--the group that perhaps is most generally representative of a potential examinee population characterized by a relatively low average levels of English proficiency--as described, for exam-ple, in the rating scale used in the present study (see Table 1, above; cf. ACTFL-scale descriptions in Appendix A).

Findings

Table 2 shows summary statistics for study variables, by group. Group 1 students, enrolled in an ESL program at the University of Delaware (and, by inference, recent or prospective TOEFL examinees planning to study in the U.S.), and Group 2 students (IEF Exchange), tested and rated after a semester of study in the U.S., scored substantially higher on the test variables and received

Table 2**Summary Statistics for the Groups**

Group	LC-EZY		R-EZY		ESL-EZY		Rating ^a		
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	
Total	66.5	15.6	69.4	15.1	135.9	29.1	5.9	2.2	(614)
1 Delaware	68.2	11.7	71.4	11.3	139.6	21.0	6.8	1.8	(74)
2 Exchange	83.0	8.2	81.7	7.7	164.7	14.4	8.5	1.5	(142)
3 Summer	47.0	11.1	52.8	12.0	99.8	20.4	4.7	1.7	(69)
4 ECC Japan	63.2	12.6	67.1	14.5	130.3	25.4	4.8	1.4	(329)
Japanese	62.0	13.8	65.8	14.7	127.8	26.6	5.1	1.7	(468)
Other	81.2	11.8	81.0	9.5	162.2	20.2	8.5	1.5	(131)
Age/Japanese examinees (in years)									
Age (15-19)	58.1	11.8	61.4	13.4	119.5	23.6	4.1	1.1	(96)
Age (20-24)	63.7	12.9	67.4	15.0	131.1	25.9	4.9	1.4	(168)
Age (25+)	69.2	10.2	74.8	11.1	144.0	19.1	5.4	1.4	(65)
Age/Total rated sample (in years)									
Age (15-19)	66.5	18.0	68.4	16.2	134.9	32.8	6.2	2.5	(315)
Age (20-24)	64.4	12.5	68.0	14.4	132.4	25.0	5.1	1.6	(197)
Age (25+)	69.3	10.4	74.8	10.7	144.1	19.2	5.8	1.6	(84)

Note. Individuals who did not respond to the background questionnaire on national origin and age were not included in the corresponding analyses.

^a Means indicate average levels according to the study scale.

higher average ratings than did either Group 3 students (in a Japanese summer program) or Group 4 students (in intensive ESL programs in Japan). The behavioral descriptions associated with the average scale-placement of the groups suggest not only that the groups involved differ widely in average level of English proficiency, but also indicate what type of linguistic behavior is typical for the respective groups (see Table 1, above, for descriptions of behavior by level).

Thus, for example, the typical IEF-Exchange (Group 2) student, judging from the mean rating of 8.5, was able to handle with confidence most social situations, including introductions and casual conversations about current events, and to handle high-frequency constructions quite

accurately. Judging from the mean rating of 6.8 (approaching Level 7) for Group 1, the typical University of Delaware, ESL-trainee involved could linguistically satisfy most survival needs and limited social demands, initiate and sustain a general conversation, and so on. Means for Groups 3 and 4 (composed of native-Japanese speaking students) were at approximately Level 5 on the schedule, at which ". . . in areas of immediate need or on very familiar topics, the (typical) student can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversation . . ." and so on.

Considering only the subgroups of Japanese examinees defined by age, of the younger group (ages 15 to 19 years) the mean of 4.1 on the abbreviated ACTFL-related rating schedule suggests that many probably were functioning at a level no higher than ILR-Level 0 Plus (see Appendix for detail regarding correspondence between the ACTFL scale and the ILR scale)--thus are at the lower levels of proficiency at which ESL-EZY, by design, might be expected to provide useful discrimination. Among Japanese examinees who were tested and rated in Japan, the means by age varied directly and positively with both average level of tested proficiency (means on LC-EZY and R-EZY) and average level of rated proficiency (mean rating). In Japan, secondary-level students are required to study English as a foreign language, hence average levels of proficiency should tend to increase with age/grade level. Thus, the covariation noted above may be thought of as attesting--indirectly--to the validity of both the test and the ratings.

Given the formal curriculum-based pattern of English-language acquisition, measures of English proficiency (especially ESL reading skills, which receive heavy curricular emphasis; e.g., see Saegusa, 1983: 101-102) logically can be treated as "achievement" measures, notwithstanding potential effects of differences in "experience out of school" on the development of proficiency. For university-educated examinees, six-years of pre-university EFL study will have been supplemented by two years of required university-level study, plus any supplementary instruction in English that may have been taken. Younger examinees are largely either secondary-level students or recent graduates with six or fewer years of EFL study.

However, this systematic positive covariation in means across age-groups did not obtain in the total rated sample, plausibly because the younger students in the IEF Exchange-Program sample were positively selected on variables associated with proficiency in English as a foreign language, including motivation to study in an English-medium environment, from subpopulations in which levels of acquired English proficiency may tend to vary directly with age/grade level, as for the Japanese ESL students.

General Regression Results

Selected findings from the analysis of relationships between scores on ESL-EZY and teachers' ratings in the samples are provided in Table 3, which shows zero-order correlation coefficients for LC-EZY, R-EZY, and ESL-EZY (the total score), respectively, with ratings, as well as corresponding multiple-correlation coefficients (for regression-weighted composites of the two section scores).

ESL-EZY/Rating correlations were quite strong in the total sample (e.g., $r = .74$ for ESL-EZY), reflecting, in part, the heterogeneous nature of the sample and the strong positive covariation for

Table 3**ESL-EZY/Criterion Relationships in Designated Samples**

Group	N	LC-EZY r	R-EZY r	ESL-EZY r	R [#]
Total sample	614	.72	.70	.74	.74
1 Delaware	74	.48	.52	.54	.54
2 IEF Exchange	142	.45	.52	.53	.54
3 Summer program	69	.32	.41	.41	.42
4 Japan (ECC)	329	.70	.74	.77	.77
Language					
Japanese	468	.59	.62	.64	.65
Other	131	.61	.64	.65	.66
Age					
Japan (ECC only)					
Age (15-19)	96	.71	.72	.77	.77
Age (20-24)	168	.70	.74	.77	.77
Age (25 +)	65	.52	.61	.62	.63
Total rated sample					
Age (15-19)	315	.78	.76	.80	.80
Age (20-24)	197	.66	.69	.72	.72
Age (25 +)	64	.47	.56	.56	.57

Note. See Table 2 for corresponding descriptive statistics.

[#] Multiple correlation coefficient (Rating vs. LC-EZY/R-EZY).

subsample means. Correlations were also strong in the various samples of younger examinees (e.g., $r = .77$ for ESL-EZY, in the sample of Japanese ECC students 15 to 19 years of age, and $r = .80$ in the corresponding age category when all rated examinees were included). Within Groups 1 and 2, both relatively highly selected on the test variables, observed ESL-EZY/Rating correlations were moderate ($r = .54$) and the coefficients for Group 3 were generally somewhat lower (e.g., $r = .41$ for ESL-EZY vs. ratings)—plausibly due, at least in part, to the fact that ratings for this group were based on shorter periods of observation than were the ratings of the other groups (see discussion of rating conditions, above).

The multiple-correlation coefficients for the regression-weighted composite of LC-EZY and R-EZY were in almost every instance not appreciably larger than the zero-order coefficient for the total ESL-EZY score (the simple sum of the two section scores) with the rating. The patterning of zero-order coefficients for LC-EZY and R-EZY, respectively, with the rating was inconsistent with a priori expectation--that is, within 11 of 12 subgroups for which analyses were conducted, the observed zero-order test/rating coefficient for LC-EZY was smaller than the corresponding coefficient involving R-EZY, contrary to the general expectation that measures of listening comprehension and speaking ability should tend to be more closely associated than are measures of reading comprehension and speaking ability. This outcome is considered in detail later on, in discussing findings.

Further Analysis of ECC Data

The relatively large sample of students in Japanese ESL programs (ECC) appears to be most representative of a population at the lower levels of developed ESL proficiency for which a test such as ESL-EZY was designed to be useful, judging from data provided in Table 4, which shows distributions of teachers' ratings (in percent) by level of performance on ESL-EZY, for the total ECC sample, and in Table 5 which shows patterns of average performance on the study variables, and intercorrelations of the variables, for ECC students in three subgroups by age.

The distributions of teachers' ratings by ESL-EZY score levels in Table 4 suggest that both ESL-EZY and teachers' ratings were sensitive to differences in proficiency at lower levels of rated proficiency, consistent with their common objective. More generally, the data in Table 5 indicate the typical level and range of rated proficiency, as defined in Table 1, associated with specified levels of performance on ESL-EZY in the sample. On balance, observed correlations between ESL-EZY and teachers' ratings, shown in Table 5, were generally relatively as strong, or stronger, in the ECC sample than in the total sample, and this was especially true for the two younger, lower-performing age groups.

DISCUSSION

These findings attest to the concurrent validity of ESL-EZY scores with respect to ESL teachers' ratings of oral English proficiency, as defined by the rating schedule used, in diverse settings in the United States and in Japan. The findings indicate that teachers were able to distinguish levels of proficiency at the low end of the ACTFL-based rating schedule employed in the study--and by inference, at corresponding levels on the basic ACTFL scale. They suggest as well that ESL-EZY scores were also sensitive to those lower-level differences.

In evaluating the positive, validity-related findings of this study, consider that teachers' ratings of proficiency were based only on nominally ACTFL-parallel excerpts, not the full ACTFL generic descriptions of proficiency levels. Moreover, the teachers involved were not trained to use the rating schedule employed, and both native English-speaking teachers in the U.S., and nonnative English-speaking ESL teachers in Japan were involved in the ratings. The strong correlations reported above suggest that the modified level-descriptions in the teachers' rating schedule (see

Table 4. Distribution of Teachers' Ratings of Oral English Proficiency, by ESL-EZY Total Score Categories (in percent) for Japanese ECC Students

ESL-EZY	Rating (see Table 1, above, for description)								Total	
	1	2	3	4	5	6	7	8		9
	(N)									
165+					3	50	12	27		(26)
145-160			1	2	47	28	9	13		(98)
125-140			3	39	47	9	1			(87)
105-120		1	21	54	20	3				(70)
85-100		12	54	27	6					(33)
< 85	7	40	33	20						(15)
Total	1	11	42	86	106	50	13	20		(329)
(%)	*	3	13	26	32	15	4	6		

Note. These data are for ECC students tested between July 1988 and September 1988, by native Japanese-speaking ESL teachers using the rating schedule shown in Exhibit A (cf., Exhibit A and Appendix A). Percentages may not total 100 percent due to rounding; "*" = < .5%.

Table 5. Means, Standard Deviations, and Intercorrelations of Variables, by Age-Group: ECC Examinees

Variable	ESL-EZY r	LC-EZY r	R-EZY r	Mean	S.D.
Age 15-19 years (N=96)					
Rating	.77	.71	.72	4.1	1.1
ESL-EZY	<u>.92</u>	<u>.94</u>		119.5	23.6
LC-EZY			.74	58.1	11.8
R-EZY				61.4	13.4
Age 20-24 years (N=168)					
Rating	.77	.70	.74	4.9	1.4
ESL-EZY	<u>.93</u>	<u>.95</u>		131.1	25.9
LC-EZY			.78	63.7	12.9
R-EZY				67.4	15.0
Age 25 years or older (N=65)					
Rating	.62	.52	.61	5.4	1.4
ESL-EZY	<u>.90</u>	<u>.91</u>		144.0	19.1
LC-EZY			.64	69.2	10.2
R-EZY				74.8	11.1

Note. Underlined coefficients are spuriously high due to self-correlation.

Table 1, above) may have captured much of the functional descriptive power inherent in the full generic descriptions (see Appendix A, Exhibit A.1) whose basic validity is attested indirectly by positive findings involving the modified, study schedule.

Study findings indicate that the native Japanese-speaking ESL teachers were able to classify their students according to the behavioral descriptions in the schedule as validly and, by inference, as reliably as did their native English-speaking counterparts. Further research designed specifically to examine similarities and differences in validity and reliability for ratings by native- and nonnative-speakers of English appears to be warranted. It would be useful to include comparable sets of assessments for the four basic macroskills.

The present study was not designed specifically to address questions of discriminant validity. At the same time, it is noteworthy that within each of the four subgroups, ESL teachers' ratings varied more closely with reading comprehension ability (R-EZY) than with listening comprehension (LC-EZY).^d The opposite pattern of validity would be expected for a measure of listening comprehension and a direct measure of "oral English proficiency;" and such a pattern has been observed in research involving the TOEIC test itself (e.g., Wilson, 1989) in which rated performance in Language Proficiency Interviews has been found to vary more systematically with TOEIC listening comprehension than with TOEIC reading comprehension.

In evaluating this finding, it is useful to keep in mind that (a) Language Proficiency Interviews are exclusively focused on eliciting and evaluating only "oral English proficiency," as reflected in the ability to produce (and, indirectly, to comprehend) utterances in English, but (b) the ESL teachers' ratings, in contrast, were based on naturalistic observation of linguistic behavior in classroom settings. Accordingly, as suggested at the outset, it is plausible that the teachers' ratings of "oral English proficiency" may have reflected teachers' holistic impressions of proficiency, shaped by daily, incidental observation of diverse aspects of students' overall linguistic performance, including their writing and reading skills, for example.

In evaluating the foregoing possibility, it is pertinent to note that similar lack of discriminant validity for generally similar types of measures was reported by Boldt, Larsen-Freeman, Reed, and Courtney (1992) in an investigation of relationships between ESL teachers' ratings of listening, writing, and reading skills based on the corresponding ACTFL schedules, on the one hand, and TOEFL section scores for listening comprehension, structure and written expression, and reading comprehension, on the other. The pattern of findings was similar to that in the present study: the level of correlation between the indirect measures (TOEFL section scores) and direct measures (teachers' ratings) under consideration was relatively strong, but the pattern of correlations for logically corresponding indirect and direct measures was not consistent with expectation.^e

As in the present study, the teachers involved in Boldt et al. (1992), were not trained in the use of the respective rating schedules, and ratings were based on naturalistically observed samples of linguistic behavior--that is, formally defined, standard samples of behavior from clearly delineated skill domains (e.g., writing samples, interviews, standard oral reading exercises) were not used.

Some Research Implications

Further research is needed to assess the relative validity of section scores on a test such as ESL-EZY for predicting ratings based on clearly delineated behavior samples (e.g., interviews for assessing "oral English proficiency"; standard writing samples). It is important, as well, to ascertain the general degree of agreement, with respect to both rank order and placement, between ratings by ESL teachers (not specially trained in rating procedures), using either abbreviated, nominally ACTFL-parallel schedules or the generic schedules, and ratings by individuals trained in ACTFL procedures.

In this study, ratings were rendered by both native English-speaking and native Japanese-speaking ESL teachers. Questions as to the comparability of judgments of nonnative speakers and native speakers of a target language regarding nonnative speakers' communicative ability, as noted at the outset, were not directly at issue in this study.

By inference from the relatively strong correlations between ratings and ESL-EZY scores, native Japanese-speaking English teachers were able to rate their students reliably and validly. This tends to confirm and extend evidence that nonnative-speaking and native-speaking teachers of a target language tend to agree, at least with respect to rank order, in rating the proficiency of non-native speakers of that language (e.g., Ingram, 1985, pp. 254-255). It would be useful to conduct research designed to extend this line of inquiry by assessing, for example, the degree of agreement between ratings of speech samples by nonnative English-speaking ESL teachers and by native English-speaking ESL specialists, respectively.

Also pertinent in this context are theoretically and pragmatically important questions regarding the reliability and validity of judgments of second-language users' communicative competence by native speakers who are "linguistically naive", as compared to the judgments of language specialists (e.g., ESL teachers, trained interviewer/raters, and so on). There is empirical evidence suggesting strongly that linguistically expert judges and linguistically naive judges are likely to agree regarding the relative comprehensibility, intelligibility and other properties of non-native-speakers' utterances (e.g., Hadden, 1991; Clark and Swinton, 1980; Powers and Stansfield, 1985, 1989); Powers, Schedl, Wilson-Leung and Butler, 1999; Dondonoli and Henning, 1990).^f

According to this evidence, for example, taped protocols of ESL utterances that are elicited and rated by professionals in formal Language Proficiency Interviews tend to be similarly perceived--in terms of relative comprehensibility, communicative content and import, judged adequacy for specified purposes, and so on--by naive native-speakers (e.g., Hadden, 1991; Powers and Stansfield, 1985, 1989; Dondonoli and Henning, 1990). Moreover, ESL speakers whose utterances are rated at relatively high (or low) levels based on samples of speech elicited in Language Proficiency Interviews tend to produce, in pragmatic, real-life (classroom teaching) settings, ESL utterances that are similarly rated as relatively higher or lower in terms of degree of interference with understanding--for example, by linguistically naive (student) receivers (Clark and Swinton, 1980).

In sum, it appears that ESL speakers who are judged to exhibit relatively high levels of ESL speaking proficiency in the interview situation or in taped speech samples may tend to produce ESL

utterances in real-life settings that are perceived by linguistically naive receivers to be relatively comprehensible, noninterfering with understanding, acceptable for specific communicative purposes, and so on.⁸ This is of considerable importance generally, because judgments by linguistically naive receivers constitute the ultimate, pragmatic test of second-language users' "communicative competence" (level of functional ability to engage in valid exchanges of meaning involving the spoken target language, for example) in real-life classroom and workplace settings. The studies cited above involved native English-speakers' perceptions of the comprehensibility or adequacy of ESL speakers' communicative skills. In the worldwide ESL testing context, it is also important to extend this line of inquiry to include assessment of the "effectiveness" of ESL communication in the growing number of situations in which both speakers and receivers are ESL users/learners.

Extending Research to School Settings

Within the Japanese population, the importance attached to proficiency in English is indicated by the fact that the study of English is a required subject beginning with middle school and continuing through the first two years of university education. Studies concerned with the documentation of EFL instructional outcomes in samples of students and/or workers can provide useful information for purposes of evaluation and planning (see, for example, Saegusa, 1983, 1985, 1989).

Findings of this study for the sample of students in ESL programs (ECC) in Japan (Group 4) suggest that the younger students in this sample tend to be functioning at the lower end of the basic ILR scale. However, younger individuals enrolled in intensive ESL programs are not necessarily representative of their nonenrolled age/grade counterparts with respect to acquired levels of ESL proficiency, motivation, or other factors that may affect performance on proficiency measures.

Assessments involving representative samples from middle and secondary school populations, and samples of university students are needed to document and evaluate rates and patterns of proficiency development in English as a foreign language in general student populations under usual conditions of instruction. Moreover, in a context in which all students--not simply those who elect to do so--study English (or some other language) as a foreign language over several years, it is particularly propitious from both pragmatic and theoretical perspectives to conduct research designed to identify cognitive and noncognitive attributes of students that are most predictive of individual differences in target-language proficiency after designated periods of study (see, for example, Lett and O'Mara, 1990; Sparks, Ganshow and Patton, 1995; Sparks, Ganshow, Patton, Artzer, Siebenbar and Plageman, 1997).^h

A powerful model--conceptually and pragmatically speaking--for assessing functional attainments in defined populations of foreign-language learners/users is represented by the work of Carroll (e.g., Carroll, 1967a, 1967b) in his benchmark national survey of the attainment of foreign language majors toward the end of the senior year in colleges and universities in the United States. Among other things, the Carroll model draws on empirical, statistical assessments of relationships between indirect and direct measures in samples from a general population to calibrate standard-score scales used in indirect, multiple-choice tests to quasi-absolute, behaviorally defined proficiency scales, such as the ACTFL scale shown in Appendix A, for example.ⁱ In any event, assessments of English proficiency of secondary school students could be designed to provide

evidence regarding functional levels attained by the end of each year of study, and information regarding typical rates and patterns of development of integrated linguistic skills under typical instructional conditions. Test-based proficiency assessments would be useful for monitoring ESL skills development and identifying unusually effective instructional settings and related instructional practices.

REFERENCES

- Alderson, C. J., Krahnke, K. J., & Stansfield, C. W. Eds. (1987). Reviews of English Language Proficiency Tests. Washington, D.C.: Teachers of English to Speakers of Other Languages.
- Angelis, P. J., Swinton, S. S., and Cowell, W. R. (1979). The performance of nonnative speakers of English on TOEFL and verbal aptitude tests (TOEFL Research Report No. 3). Princeton, NJ: Educational Testing Service.
- Boldt, R. F., Larsen-Freeman, D., Reed, M. S., Courtney, R. G. (1992). Distributions of ACTFL ratings by TOEFL score ranges (TOEFL Research Report No. 41). Princeton, NJ: Educational Testing Service.
- Breland, H. M. (1977). A study of college English placement and the Test of Standard Written English (College Board RDR-76-77, No. 4, and ETS Project Report, PR-77-1). Princeton, NJ: Educational Testing Service.
- Brumfit, C., (Ed.). (1982). English for International Communication. N. Y.: Pergamon Press.
- Carlson, S., Bridgeman, B., Camp, R., and Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English (TOEFL Research Report No. 19). Princeton, NJ: Educational Testing Service.
- Carroll, J. B. (1967a). The foreign language attainments of language majors in the senior year: A survey conducted in United States colleges and universities (Final Report, Contract OE-4-14-048). Cambridge, MA: Harvard University Graduate School of Education.
- Carroll, J. B. (1967b). Foreign language proficiency levels attained by language majors near graduation from college, Foreign Language Annals, 1(2), 131-151.
- Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones and B. Spolsky Eds., Testing Language Proficiency (pp. 29-44). Arlington, VA: Center for Applied Linguistics.
- Clark, J. L. D., and Swinton, S. S. (1979). An exploration of speaking proficiency measures in the TOEFL context (TOEFL Research Report No. 4). Princeton, NJ: Educational Testing Service.

- Clark, J. L. D., and Swinton, S. S. (1980). *The Test of Spoken English as a test of communicative ability in English-medium instructional settings (TOEFL Research Report No. 7)*. Princeton, NJ: Educational Testing Service.
- Dondonoli, P., and Henning G. (1990). An investigation of the construct validity of the ACTFL Proficiency Guidelines and Oral Interview Procedure, *Foreign Language Annals*, 23, 11-22.
- Echternacht, G. (1970). TOEFL score ranges and theme writing samples (Statistical Report SR-70-8). Princeton, NJ: Educational Testing Service.
- Educational Testing Service (1982). *ETS Oral Proficiency Testing Manual*. Princeton, NJ: Author.
- Educational Testing Service (1991). *TOEFL Test and Score Manual*. Princeton, NJ: Author.
- Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication, *Language Learning*, 41, 1-24.
- Hale, G. A. (1986). An overview of research related to the TOEFL. In C. W. Stansfield (Ed.), *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference (TOEFL Research Report No. 21)*: pp. 10-16): Princeton, NJ: Educational Testing Service.
- Hilton, T. L., Grandy, J., Kline, R. G., & Liskin-Gasparro, J. E. (1985). *The oral language proficiency of teachers in the United States in the 1980's--An empirical study*. Princeton, NJ: Educational Testing Service.
- Hyltenstam, K. and Pienemann, M. (Eds.). (1985). *Modelling and Assessing Second Language Acquisition* (pp. 277-282). San Diego, CA: College Hill Press.
- Ingram, D. E. (1985). *Assessing proficiency: An overview of selected aspects of testing*. In K. Hyltenstam and M. Pienemann Eds., *Modelling and Assessing Second Language Acquisition* (pp. 277-282). San Diego, CA: College Hill Press.
- Lett, J. A., Jr., and O'Mara, F. E. (1990). Predictors of success in an intensive foreign language learning context: Correlates of language learning at the Defense Language Institute Foreign Language Center, In T. S. Parry and C. Stansfield (Eds.), *Language Aptitude Reconsidered* (pp. 223-260). Englewood Cliffs, NJ: Englewood Cliffs, NJ: Prentice Hall Regents.
- Lowe, P. (1987). Interagency language roundtable oral proficiency interview. In C. J. Alderson, K. J. Krahnke, & C. W. Stansfield Eds., *Reviews of English Language Proficiency Tests* (pp. 43-47). Washington, D.C.: Teachers of English to Speakers of Other Languages.
- Lowe, P. L., Jr. and Stansfield, C. W. (Eds.)(1988a). *Second Language Proficiency Assessment: Current Issues*. Englewood Cliffs, NJ: Prentice Hall Regents.

- Oller, J. W. Ed. (1983). Issues in Language Testing Research. Rowley, MA: Newbury House.
- Perkins, K. (1987). Test of English for International Communication. In C. J. Alderson, K. J. Krahnke, and C. W. Stansfield (Eds.), Reviews of English Language Proficiency Tests (pp. 81-83). Washington, D. C.: Teachers of English to Speakers of Other Languages.
- Pike, L. W. (1979). An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report No. 2 and ETS RR-79-6). Princeton, NJ: Educational Testing Service.
- Powers, D. E. (1980). The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language (TOEFL Research Report No. 5). Princeton, NJ: Educational Testing Service.
- Powers, D. E., and Stansfield C. W. (1985). Testing the oral English proficiency of foreign nursing graduates, The ESP Journal, 4, 21-35.
- Powers, D. E., and Stansfield, C. W. (1989). Communicative ability in three health professions. In H. Coleman (Ed.), Working with Language: A Multidisciplinary Consideration of Language Use in Work Contexts (pp. 341-356). New York: Mouton de Gruyter.
- Powers, D. E., Schedl, M. A., Wilson-Leung, S., and Butler, F. A.(1999). Validating the Revised Test of Spoken English against a criterion of communicative success (TOEFL Research Report No. 63 and ETS RR-99-5). Princeton, NJ: Educational Testing Service.
- Saegusa, Y. (1989). Japanese company workers' English proficiency, WASEDA Studies in Human Sciences, 2, 1-12.
- Saegusa, Y. (1985). Prediction of English proficiency progress. Musashino English and American Literature, 18, 65-185.
- Saegusa, Y. (1983). Japanese college students' reading proficiency in English. Musashino English and American Literature, 16, 99-117.
- Sharon, A. T. (1972). English proficiency, verbal aptitude, and foreign student success in American graduate schools. Educational and Psychological Measurement, 32, 425-431.
- Sparks, R., Ganshow, L., Patton, J., Artzer, M., Siebenhar, D., and Plageman, M. (1997). Prediction of foreign language proficiency. Journal of Educational Psychology, 89, 549-556.
- Sparks, R., Ganshow, L., and Patton, J. (1995). Prediction of performance in first-year language courses: Connections between native and foreign language learning. Journal of Educational Psychology, 87, 638-655.
- The Chauncey Group International (1996). TOEIC: Report on Test-Takers Worldwide. Princeton, NJ: Author.

- The Chauncey Group International (1999). TOEIC User Guide. Princeton, NJ: Author.
- Wilson, K. M. (1989). Enhancing the interpretation of a norm-referenced second-language test through criterion-referencing: A research assessment of experience in the TOEIC testing context (TOEIC Research Report No. 1 and ETS RR-89-39). Princeton, NJ: Educational Testing Service.
- Wilson, K. M. (1986). The relationship of GRE General Test scores to first-year grades for foreign graduate students (GRE Board Professional Report GREB No. 82-11P and ETS-RR-86-44). Princeton, NJ: Educational Testing Service.
- Wilson, K. M. (1985). Factors affecting GMAT predictive validity for foreign MBA students: An exploratory study (ETS-RR-85-17). Princeton, NJ: Educational Testing Service.
- Wilson, K. M. (1982). GMAT and GRE Aptitude Test performance in relation to primary language and scores on the TOEFL (TOEFL Research Report No. 12). Princeton, NJ: Educational Testing Service.
- Woodford, P. E. (1982). The Test of English for International Communication (TOEIC). In C. Brumfit (Ed.), English for International Communication (pp. 61-72). New York: Pergamon Press.
- Yule, G., and Hoffman, P. (1991). Predicting success for international teaching assistants in a U.S. university. TESOL Quarterly, 24, 227-241.

ENDNOTES

- a. These TOEIC findings are consistent with general evidence that various aspects of English proficiency are relatively closely intercorrelated: it appears that coefficients centering around .70 can be expected between direct and indirect measures of basic ESL macroskills (writing, speaking, listening, reading) in representative samples of educated ESL users/learners, such as those who take the TOEIC in corporate settings, or the TOEFL in academic settings (see, for example, Hale, 1986; Carlson, Bridgeman, Camp, and Waanders, 1985; Oller, 1983; Swinton and Clark, 1979, 1980; Pike, 1979; Echternacht, 1970; Carroll, 1967a, 1967b). This level of correlation (centering around .7) also tends to obtain between scores on the TOEFL and scores on the verbal sections of standard undergraduate- and graduate-level admission tests (e.g., SAT, GMAT, GRE), widely used in the United States, in both unselected samples (e.g., Wilson, 1982; Powers, 1980; Angelis, Swinton, and Cowell, 1979) and highly selected samples of enrolled graduate students (e.g., Yule and Hoffman, 1991; Wilson, 1986, 1985; Sharon, 1972).
- b. Materials forwarded by ETS staff to the Group 1 site (University of Delaware), included a copy of the generic ACTFL descriptions (Appendix A); the ESL teachers who rated the Japanese students in Group 3 (summer ESL programs) received a complete manual for interviewing/rating examinees (derived from ETS, 1982) as well as the generic descriptions. The materials were forwarded simply as potentially useful information for the teacher/raters in these particular sites, not as part of the basic study design. The study was not designed to assess possible effects on ratings of differences in the type and amount of additional information available to the EFL/ESL teachers involved.
- c. For present purposes, regarding the reliability issue it is useful to recall certain of Thorndike's (1949) observations regarding the relative importance of validity and reliability in criterion variables. For example: "The quality designated as 'relevance to the ultimate goal' is the prime essential in a criterion measure. A criterion measure is relevant as far as the knowledge, skills, and basic aptitudes required for success on it are the same as those required for performance of the ultimate task. . . . A necessary but not a sufficient condition . . . is that the criterion measure have some reliability. . . . Low reliability in a criterion measure merely attenuates all its relationships with other measures" (pp. 126-127). Teachers' ratings have considerable general face validity, and the relatively high levels of correlation observed in the present study suggest that the ratings were relatively reliably rendered as well. Reliability coefficients, if available, would permit estimation of correlations between ESL-EZY scores and teachers' ratings assuming absence of measurement error in both measures, for example--an issue that is of theoretical significance but not at issue in the present study.
- d. The pattern of correlations between ESL-EZY sections and teachers' ratings in the total sample was consistent with previous findings involving the corresponding TOEIC sections and LPI ratings, even though this was not the case for the within-group pattern--that is, as shown in Table 4 in the text, the LC-EZY/Rating correlation actually was slightly higher than the R-EZY/Rating correlation ($r = .72$ vs. $r = .70$). In evaluating the foregoing, recall (from Table 1, text) that a wide range of proficiency was represented in the total sample, due especially to the presence of the group of IEF exchange-students who were highly selected, directly or indirectly, on "oral English proficiency" (for example, the average rating was about 1.2 standard deviations higher than

that for the total sample). By inference, this group was also more highly selected on listening comprehension than on reading ability, averaging more than 1.0 standard deviation above the total sample average on the former but only about .8 standard deviation higher on the latter.

e. Boldt, Larsen-Freeman, Camp, and Levin (1992) investigated relationships between (a) ratings of listening, writing, and reading, respectively, rendered by ESL teachers according to the corresponding, formal ACTFL schedules and (b) TOEFL section scores for Listening Comprehension (LC), Structure and Written Expression (SWE), and Reading Comprehension (RC). Correlations in the .60 to .70 range were reported--generally comparable to the level reported for ESL-EZY and teachers' ratings of oral English proficiency in the present study. However, ACTFL ratings of listening, reading, and writing, respectively, did not covary more systematically with scores on the logically corresponding TOEFL sections than with scores on other sections. For example, TOEFL Listening Comprehension was not a better predictor of the ACTFL listening rating than were scores on TOEFL Reading Comprehension and Vocabulary, or Structure and Written Expression. As in the present study, the teachers involved in Boldt et al. were not trained in the use of the respective rating schedules, but perhaps of more significance in explaining study outcomes is the fact that in Boldt et al., as in the present study, the ratings were based on naturalistically observed samples of linguistic behavior--that is, not on standard samples of behavior from clearly delineated skill domains (e.g., writing samples; standard reading and/or dictation exercises). And, in both studies the pattern of findings was similar: the level of correlation between the indirect and direct measures under consideration was relatively strong, but the pattern of observed correlation for logically corresponding indirect and direct measures was not consistent with expectation. In any event, these circumstances suggest that lack of criterion specificity, common to both studies, limits inferences from the study findings regarding discriminant validity properties of the ESL proficiency measures involved. At the same time, outcomes of both studies extend evidence indicating that measures of basic ESL macroskills tend to be relatively highly intercorrelated; and the outcomes attest to the concurrent validity properties of the measures involved as well.

f. Hadden (1991) analyzed the reactions of ESL teachers and linguistically naive students (undergraduate students in education courses) to videotaped protocols for several ESL speakers. Based on a factor analysis of reactions by these two groups of "receivers," Hadden reported, in part, as follows: "[There was] notable similarity in ESL teacher and nonteacher perceptions of second-language communication Moreover the teacher and nonteacher ratings [on four of the five dimensions identified]--comprehensibility, social acceptability, personality, and body language--did not differ significantly. . . . Only in regard to linguistic ability, that aspect of second-language communication typically of /major concern to the language teacher, did teacher and nonteacher perceptions differ significantly" (pp. 17-18). Clark and Swinton (1980: Table 10) analyzed ratings by linguistically naive students in several U. S. universities, of nonnative-English speaking teaching assistants' . . . "overall ability to communicate (in English)." Students rated the teaching assistants, based on experience in classroom and laboratory settings, according to the extent to which "the instructor's overall ability to communicate" interfered with understanding, on a scale ranging from "did not interfere" = 1 to "interfered completely" = 5. Students' ratings on this negatively scaled criterion measure were relatively highly correlated ($r = -.72$) with previously obtained ILR/FSI-scaled ratings of oral English proficiency, professionally assessed especially for the study through the LPI procedure.

Powers and Stansfield (1985, 1989) had linguistically naïve native-English speakers judge the "acceptability" of representative Test of Spoken English (TSE) protocols (recorded speech samples) of foreign nursing graduates (in the 1985 study), and of nonnative-English speaking pharmacists, physicians, and veterinarians (in the 1989 study). Judges were native-English speaking members of the respective professions, and consumers, respectively. Powers and Stansfield (1989: p. 2) reported relationships between ratings of acceptability (for specified purposes) and TSE scores, as follows: "The median product-moment correlations (over ratings and trials) between judges' ratings and TSE scores were .59, .72., .62, and .68 for pharmacists, physicians, veterinarians and consumers, respectively. Similarly strong relationships were reported in the study involving nurses only. In a later study (Powers, Schedl, Wilson-Leung and Butler, 1999)--also designed to determine the degree to which official TSE scores are predictive of linguistically naïve listeners' ability to understand the messages conveyed by TSE examinees--it was reported as follows: "Analyses revealed a strong association between TSE score levels and the judgments reactions, and understanding of listeners. This finding applied to all TSE tasks and to nearly all of the several different kinds of evaluations made by listeners" (p. i). The study sample involved "undergraduate students . . . selected as 'evaluators' because they, more than most other groups, are likely to interact with TSE examinees, many of whom become teaching assistants."

Dondonoli and Henning (1990: p. 20), among other things, had linguistically naive native-speakers rank-order taped protocols representing the full range of ACTFL-defined proficiency levels in English as a second language and French as a second language, respectively, based on ratings by ACTFL-trained raters. According to the study report: "For each set of tapes, two untrained native speakers were asked independently to rank the tapes in order of increasing ability using whatever criteria they chose. For English, [Spearman rank] correlations ranged from 0.904 to 1.000. . . ; for French, the correlations ranged from 0.857 to 1.000 with a mean of .929" (p. 20).

Evidence such as that cited above suggests strongly that, as receivers of indirect (e.g., taped protocols) or direct oral communication by nonnative-speakers at differing levels of proficiency in a target language, both linguistically naïve raters and linguistically sophisticated raters (e.g., professionals trained in the Language Proficiency Interview or related ACTFL procedure) tend to share generally similar perceptions regarding the relative ability (rank-ordering) of the nonnative-speakers involved "to produce comprehensible utterances" in the target language. There are, of course, meaningful distinctions between the task of rank-ordering given speech samples and that of assigning them to specifically defined proficiency levels (as in the LPI procedure, for example).

g. Findings indicating agreement among diverse receivers as to the relative intelligibility of utterances of nonnative speakers of a target language as elicited in Language Proficiency Interviews or in response to recorded prompts are important from a psychometric perspective because they attest to the "authenticity" of the samples of linguistic behavior that are elicited in these types of "testing" situations.

h. Lett and O'Mara (1990) report results of systematic assessments conducted by the Defense Language Institute Foreign Language Center, of language-learning rates for native-English speaking members of the U. S. Armed Forces after periods of intensive foreign-language training in over 40 different target languages, including relationships between a variety of predictors (demographic, cognitive, aptitude and personality variables) and end-of-course proficiency measures for each of

four target-language subgroups. Sparks, Ganshow and Patton (1995) studied the predictive validity of measures of aptitude for foreign-language learning and native language verbal aptitude, respectively, along with indices of prior academic achievement, using as criteria end-of-course grades in Spanish and French, respectively, for a sample of U.S. high-school students studying these languages. In a follow-up study, Sparks, Ganshow, Patton, Artzer, Siebenhar and Plageman (1997) employed teachers' ratings and grades at the end of the second year of study as criteria.

i. Generally speaking, the findings that have been reviewed extend and confirm a growing body of empirical evidence attesting to the validity of the following general, logically related propositions. First, indirect and direct measures of ESL proficiency can be expected to be relatively strongly intercorrelated in samples of educated ESL, and other second-language users/learners (e.g., Hale, 1986; Oller, 1983; Swinton and Clark, 1979, 1980; Pike, 1979; Carroll, 1967a, 1967b; Echternacht, 1970) as Carroll (1967a, 1967b) demonstrated clearly more than thirty years ago. Second the interpretation of indirect tests can be enhanced by using familiar statistical models to link level of performance on particular indirect tests--typically distributed along arbitrarily scaled standard-score continua--to level of performance based on direct observation of language use (e.g., in conversational interviews, writing or reading) rated according to quasi-absolute behaviorally anchored scales, especially the well-established ILR scales (Carroll, 1967a, 1967b; Clark, 1975; Woodford, 1982; Hilton et al., 1985; Saegusa, 1985; Wilson, 1989), and the corresponding ILR-linked, ACTFL scales (Boldt et al., 1992).

APPENDIX A: Relationship Between ACTFL and ILR Scales

The ILR/FSI (or ILR) quasi-absolute proficiency scale was not designed to discriminate among ESL users/learners at the lower end of the developmental continuum. To provide such discrimination, the Association of College Teachers of Foreign Languages (ACTFL) and ETS collaborated to expand the lower end of the ILR-scale levels (e.g., ETS, 1982).

The following brief rationale for modifying the basic ILR scale is provided in ETS (1982, p. 2), which also constitutes a basic manual for the interview procedure:

"As the government (FSI) scale attracted interest within formal academic circles, a feeling grew among foreign language professionals in secondary and postsecondary education that the scale was not as sensitive at the lower end as it should be. The ILR scale . . . covers the whole spectrum of speaking ability from Level 0 ('no functional ability') to Level 5 ('ability equivalent to that of an educated native speaker of the language'). In the assessment of the speaking ability of high school and college students, the full range of the scale is rarely used. . . (It thus) seems safe to assume that the ILR levels of most interest to high school and college foreign language teachers will be levels 0 and 1."

Exhibit A.1 shows full generic descriptions for the ACTFL Oral Proficiency Interview scale. Exhibit A.2 provides perspective regarding the relationship between the ACTFL modification and the basic ILR scale.

Exhibit A.2 shows the 11 levels that make up the basic ILR scale, and designations denoting each of 10 levels on the ACTFL Oral Proficiency Interview (OPI) scale. Underlining indicates that the generic behavioral description associated with the designated ACTFL level is the same as that associated with the corresponding underlined ILR level.

The ACTFL modification is seen to differ from the basic ILR scale only in the following respects:

(a) It provides for finer discrimination at the very low end of the developmental continuum spanned by the ILR scale, by adding descriptions for ratings below ILR Level 0 Plus but above Level 0, and within ILR Level 1.

(b) It classifies all levels of proficiency ratable at or above ILR Level 3, as "Superior."

Thus, the ACTFL/ETS scale (or ACTFL scale, as it has come to be known) is embedded historically and functionally in the well-established ILR scale.

The potential usefulness of the ACTFL modification of the ILR scale as a basis for rating oral English proficiency at lower levels has not been directly assessed in the TOEIC testing context, primarily because TOEIC-sponsored operational programs of direct assessment (involving the LPI procedure and corresponding ILR-referenced ratings of speaking proficiency) have been expressly restricted so as to involve only TOEIC examinees expected to exhibit levels of oral English proficiency above those of greatest interest for ACTFL-related assessment.

EXHIBIT A.2. ACTFL PROFICIENCY GUIDELINES Page 1 of 3 pages

Generic Descriptions-Speaking

Novice The Novice level is characterized by the ability to communicate minimally with learned material.

Novice-Low Oral production consists of isolated words and perhaps a few high-frequency phrases. Essentially no functional communicative ability.

Novice-Mid Oral production continues to consist of isolated words and learned phrases within very predictable areas of need, although quality is increased. Vocabulary is sufficient only for handling simple, elementary needs and expressing basic courtesies. Utterances rarely consist of more than two or three words and show frequent long pauses and repetition of interlocutor's words. Speaker may have some difficulty producing even the simplest utterances. Some Novice-Mid speakers will be understood only with great difficulty.

Novice-High Able to satisfy partially the requirements of basic communicative exchanges by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements. Can ask questions or make statements involving learned material. Shows signs of spontaneity although this falls short of real autonomy of expression. Speech continues to consist of learned utterances rather than of personalized, situationally adapted ones. Vocabulary centers on areas such as basic objects, places, and most common kinship terms. Pronunciation may still be strongly influenced by first language. Errors are frequent and, in spite of repetition, some Novice-High speakers will have difficulty being understood even by sympathetic interlocutors.

Intermediate The Intermediate level is characterized by the speaker's ability to:

-create with the language by combining and recombining learned elements, though primarily in a reactive mode;

-initiate, minimally sustain, and close in a simple way basic communicative tasks; and, ask and answer questions.

Intermediate-Low Able to handle successfully a limited number of interactive, task-oriented and social situations. Can ask and answer questions, initiate and respond to simple statements and maintain face-to-face conversation, although in a highly restricted manner and with much linguistic inaccuracy. Within these limitations, can perform such tasks as introducing self, ordering a meal, asking directions, and making purchases. Vocabulary is adequate to express only the most elementary needs. Strong interference from native language may occur. Misunderstandings frequently arise, but with repetition, the Intermediate-Low speaker can generally be understood by sympathetic interlocutors.

Intermediate-Mid Able to handle successfully a variety of uncomplicated, basic and communicative tasks and social situations. Can talk simply about self and family members. Can ask and answer questions and participate in simple conversations on topics beyond most immediate needs; e.g., personal history and leisure time activities. Utterance length increases slightly, but speech may continue to be characterized by frequent long pauses, since the smooth incorporation of even basic conversational strategies is often hindered as the speaker struggles to create appropriate language forms. Pronunciation may continue to be strongly influenced by first language and fluency may still be strained. Although misunderstandings still arise, the Intermediate-Mid speaker can generally be understood by sympathetic interlocutors.

Intermediate-High Able to handle successfully most uncomplicated communicative tasks and social situations. Can initiate, sustain, and close a general conversation with a number of strategies appropriate to a range of circumstances and topics, but errors are evident. Limited vocabulary still necessitates hesitation and may bring about slightly unexpected circumlocution. There is emerging evidence of connected discourse, particularly for simple narration and/or description. The Intermediate-High speaker can generally be understood even by interlocutors not accustomed to dealing with speakers at this level, but repetition may still be required.

Advanced The Advanced level is characterized by the speaker's ability to:

- converse in a clearly participatory fashion;
- initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events;
- satisfy the requirements of school and work situations; and
- narrate and describe with paragraph-length connected discourse.

Advanced Able to satisfy the requirements of everyday situations and routine school and work requirements. Can handle with confidence but not with facility complicated tasks and social situations, such as elaborating, complaining, and apologizing. Can narrate and describe with some details, linking sentences together smoothly. Can communicate facts and talk casually about topics of current public and personal interest, using general vocabulary. Shortcomings can often be smoothed over by communicative strategies, such as pause fillers, stalling devices, and different rates of speech. Circumlocution which arises from vocabulary or syntactic limitations very often is quite successful, though some groping for words may still be evident. The Advanced-level speaker can be understood without difficulty by native interlocutors.

Advanced-Plus Able to satisfy the requirements of a broad variety of everyday, school, and work situations. Can discuss concrete topics relating to particular interests and special fields of competence. There is emerging evidence of ability to support opinions, explain in detail, and

hypothesize. The Advanced-Plus speaker often shows a well developed ability to compensate for an imperfect grasp of some forms with confident use of communicative strategies, such as paraphrasing and circumlocution. Differentiated vocabulary and intonation are effectively used to communicate fine shades of meaning. The Advanced-Plus speaker often shows remarkable fluency and ease of speech but under the demands of Superior-level, complex tasks, language may break down or prove inadequate.

SUPERIOR The Superior-level is characterized by the speaker's ability to:

-participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics; and

-support opinions and hypothesize using native-like discourse strategies.

Superior Able to speak the language with sufficient accuracy to participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics. Can discuss special fields of competence and interest with ease. Can support opinions and interests with ease. Can support opinions and hypothesize, but may not be able to tailor language to audience or discuss in depth highly abstract or unfamiliar topics. Usually the Superior level speaker is only partially familiar with regional or other dialectical variants. Superior level speaker commands a wide variety of interactive strategies and shows good awareness of discourse strategies. The latter involves the ability to distinguish main ideas from supporting information through syntactic, lexical and supra-segmental features (pitch, stress, intonation). Sporadic errors may occur, particularly in low-frequency structures and some complex high-frequency structures more common to formal writing, but no patterns of error are evident. Errors do not disturb the native speaker or interfere with communication.

EXHIBIT A.2 Outline of Basic ILR/FSI and ACTFL/ETS Levels ^a

Level	ILR scale	ACTFL/ETS Scale
0	<u>No proficiency</u>	<u>No proficiency</u>
		Novice-Low Novice-Mid <u>Novice-High</u>
0+	<u>Not labeled</u>	
1	<u>Elementary proficiency</u>	<u>Intermediate-Low</u> Intermediate-Mid Intermediate-High
1+	Not labeled	
2	<u>Limited working proficiency</u>	<u>Advanced</u>
2+	<u>Not labeled</u>	<u>Advanced Plus</u>
3	Minimum professional proficiency	Superior ^b
3+	Not labeled	
4	Full professional proficiency	
4+	Not labeled	
5	Native or bilingual proficiency	

Note. Underlining indicates levels that are common to both scales.

^a See ETS (1982) for a full description of the respective levels and historical perspective on steps that led to the development of the ACTFL/ETS scale. The ACTFL/ETS levels are directly anchored within the ILR/FSI scale. Underlining indicates that the level descriptions are common to both scales.

^b All levels higher than Advanced Plus.

APPENDIX B. Characteristics of ESL-EZY

ESL-EZY consisted of 120 multiple-choice questions. The test was divided into two sections, listening comprehension (60 questions) and reading comprehension (60 questions), as follows:

Listening Comprehension	Type of Item	N Items
Part I	One picture, four spoken sentences	10
Part II	Spoken utterances, three spoken responses	15
Part III	Short conversations, four printed answers	15
Part IV	Short talks, four printed questions and answers	20
Reading Comprehension		
Part V	Incomplete sentences, fill-in blanks	15
Part VI	Error recognition, underlined words	20
Part VII	Reading comprehension, passages	25

Note. Test takers were instructed to indicate their answers to the questions by marking the letter A, B, C, or D, on a scannable answer sheet. ESL-EZY required approximately 90 minutes to administer.

