VALIDITY OF THE SECONDARY LEVEL ENGLISH PROFICIENCY TEST AT TEMPLE UNIVERSITY-JAPAN

Kenneth M. Wilson
with the collaboration of Kimberly Graves
Temple University Japan
ABSTRACT

The study reported herein assessed levels and patterns of concurrent correlations for Listening Comprehension (LC) and Reading Comprehension (RC) scores provided by the Secondary Level English Proficiency (SLEP) test with direct measures of ESL speaking proficiency (interview rating) and writing proficiency (essay rating), respectively, and (b) the internal consistency and dimensionality of the SLEP, by analyzing intercorrelations of scores on SLEP item-type "parcels" (subsets of several items of each type included in the SLEP). Data for some 1,600 native-Japanese speakers (recent secondary-school graduates) applying for admission to Temple University-Japan (TU-J) were analyzed. The findings tended to confirm and extend previous research findings suggesting that the SLEP--which was originally developed for use with secondary-school students--also permits valid inferences regarding ESL listening- and reading-comprehension skills in postsecondary-level samples.

Key words: ESL proficiency, SLEP test, interview rating, essay rating, TWE
ACKNOWLEDGEMENTS

This report reflects the outcome of a cooperative study involving the Educational Testing Service and Temple University-Japan (TU-J). TU-J supplied all the basic study data which were generated operationally in a testing program conducted by TU-J's International English Language Program for ESL screening and placement purposes. TU-J was represented by Ms. Kimberly Graves who developed the basic data file used in the study, provided detailed descriptions of the TU-J context and assessment procedures, and contributed interpretive comments and suggestions during the course of the study and the preparation of the study report. Essential indirect support was provided by the ETS Research Division. Very helpful reviews of the draft were provided by Robert Boldt and Donald Powers. These contributions are acknowledged with appreciation.
INTRODUCTION

The Secondary Level English Proficiency (SLEP) test, developed by the Educational Testing Service (ETS), is a standardized, multiple-choice test designed to measure the listening comprehension and reading comprehension skills of nonnative-English speakers (see, for example, ETS, 1991; ETS, 1997). The SLEP, described in greater detail later, generates scaled scores for listening comprehension (LC) and reading comprehension (RC), respectively, which are summed to form a Total score.

As suggested by its title, the SLEP was originally developed for use by secondary schools in screening and/or placing nonnative-English speaking applicants, who were tested in ETS-operated centers worldwide. The latter practice was discontinued, but ETS continued to make the SLEP available to qualified users for local administration and scoring.

The SLEP is currently being used not only in secondary-school settings but also in postsecondary settings. For example, approximately one-third of respondents to a survey of SLEP users (Wilson, 1992) reported use of the SLEP with college-level students: to assess readiness to undertake English-medium academic instruction, for placement in ESL courses, for course or program evaluation, for admission screening, and so on.

Evidence of SLEP's validity as a measure of proficiency in English as a second language (ESL) was obtained in a major validation study associated with SLEP's development and introduction (Stansfield, 1984). SLEP scores and English language background information were collected for approximately 1,200 students from over 50 high schools in the United States.

SLEP scores were found to be positively related to ESL background variables (e.g., years of study of English, years in the U.S. and in the present school); average SLEP performance was significantly lower for subgroups independently classified as being lower in ESL proficiency (e.g., those in full-time or part-time bilingual programs) than for subgroups classified as being higher in ESL proficiency (e.g., ESL students engaged in full-time English-medium academic study).

The findings of the basic validation study and other studies cited in various editions of the SLEP Test Manual (e.g., ETS, 1991, pp. 29 ff.), along with reliability estimates above the .90 level, suggest that the SLEP can be expected to provide reliable and valid information regarding the listening and reading skills of ESL students in the G7-12 range. Other research findings, reviewed below, suggest that this can also be expected to be true of the SLEP when the test is used with college-level students. The present study was undertaken to obtain additional empirical evidence bearing on the validity of the SLEP when used with college-level students.

Review of Related Research

An ETS-sponsored study reported in the SLEP Test Manual (e.g., ETS, 1991) examined the relationship of SLEP scores to scores on the familiar Test of English as a Foreign Language (TOEFL) in a sample of 172 students from four intensive ESL training programs in U.S.
universities. SLEP Listening Comprehension (LC) score correlation with TOEFL LC was \( r = .74 \), somewhat higher than that of either the TOEFL Structure and Written Expression (SWE) score or the TOEFL Vocabulary and Reading Comprehension (RC) score; SLEP Reading Comprehension (RC) score correlated equally highly with TOEFL LC \( (r = .80) \) and TOEFL RC \( (r = .79) \); the correlation between the two total scores in this sample was .82, slightly lower than that between SLEP RC and TOEFL Total (.85).

The mean TOEFL score was at about the 57th percentile in the general TOEFL reference group distribution, while the SLEP total score mean was at about the 80th percentile in the corresponding SLEP reference group distribution. Thus, by inference, SLEP appears to be an "easier" test than the TOEFL. At the same time, based on the observed strength of association between the two measures, SLEP items apparently are not "too easy" to provide valid discrimination in samples such as those likely to be enrolled in college-level, intensive ESL programs in the United States.

The latter point is strengthened by findings reported by Butler (1989), who studied SLEP performance in a sample made-up of over 1,000 ESL students typical of those entering member institutions of the Los Angeles Community College District (LACCD). Based on analysis of average percent-right scores for various sections of the test, Butler reported that "no group of students 'topped out' or 'bottomed out' on either the subsections or the total (test)" (p.25); also that the relationship between SLEP scores and independently determined ESL placement levels was generally positive.

In a subsequent study in the LACCD context (Wilson, 1994), concurrent correlations centering around .60 were found between scores on a shortened version of the SLEP and rated performance on locally developed and scored writing tests. Coefficients tended to be somewhat higher for the shortened-SLEP reading score than for the corresponding SLEP listening score, suggesting "proficiency domain consistent" patterns of discriminant validity for the corresponding shortened SLEP sections.

Further evidence bearing on SLEP's concurrent and discriminant validity in college-level samples is provided by results of a study (Rudmann, 1990) of the relationship between SLEP scores and end-of-course grades in ESL courses at Irvine Valley (CA) Community College. Within-course correlations between SLEP scores and ESL-course grades averaged approximately \( r = .40 \), uncorrected for attenuation due to restriction of range on SLEP score which was the sole basis for placing students in ESL courses graded according to proficiency level. In courses emphasizing conversational (speaking) skills, correlations between SLEP LC score and grades were higher than were those between SLEP RC score and grades; average coefficients for SLEP reading were higher than those for SLEP listening in courses emphasizing vocabulary/grammar/reading.

The findings just reviewed indicate that SLEP section scores (for listening comprehension [LC] and reading [R], respectively) have been found to exhibit expected patterns of relationships with other ESL-proficiency measures or measures of academic attainment in EFL/ESL studies. SLEP scores appear to correlate positively and moderately, in a "proficiency domain consistent" pattern, with generally corresponding scores on a widely used and studied,
college-level English proficiency test (the TOEFL), and other measures including, for example, writing samples, grades in courses with emphasis on listening/speaking skills versus vocabulary/grammar/reading skills.

THE PRESENT STUDY

Previous studies have shed light on the validity of SLEP section and total scores in linguistically heterogeneous samples of college-level ESL students residing (tested) in the United States. The present study was undertaken to extend evidence bearing on the SLEP's validity as a measure of ESL listening comprehension and reading skills in college-level samples, using data for native-Japanese speaking students planning to enroll in the English-medium academic program being offered by Temple University-Japan (TU-J).

The data used in this study--provided by TU-J--were generated in comprehensive ESL-placement-testing sessions conducted by TU-J's International English Language Program (IELP). The IELP placement battery includes measures of the four basic ESL macroskills: listening and reading comprehension skills, as measured by the corresponding sections of the SLEP test, speaking proficiency measured using a locally developed direct interview procedure, and writing ability as assessed by writing samples in the form of essays.

The SLEP sections and the two direct measures have general face validity as measures of the corresponding underlying macroskills. And IELP placement decisions based on a composite of scores on the measures reportedly have substantial pragmatic validity--that is, there is low incidence of change in placement due to observed incongruity between levels of student functioning in classroom activities at initial placement levels and expected level of functioning based on the placement battery; also, both students and IELP faculty appear to be generally satisfied with results of the placement process (Graves, 1994, personal communication).

Moreover, results of local TU-J validity studies (see Table 1) indicate that the SLEP and the two direct measures typically exhibit relatively stronger postdictive relationships with grade point average (GPA) for high school courses in English as a foreign language (EFL GPA), than with the corresponding measure of overall academic performance (Academic GPA). This pattern is consistent with expectation for measures of ESL proficiency, given these GPA-criterion variables. Table 1 shows findings—provided by Graves (1994, personal communication) for the TU-J sample involved in this study-- said to be typical for such samples. Evidence such as the foregoing provides general support for the validity of the SLEP and the two direct measures as measures of ESL proficiency in the TU-J context.

1 Note also that correlations for SLEP RC with the EFL GPA criterion are slightly larger than those for SLEP LC, a pattern that tends to obtain for essay rating versus interview rating as well. This pattern is consistent with the assumption of greater EFL curricular emphasis in Japanese secondary schools (e.g., Saegusa, 1983; Saegusa, 1985) on development of reading skills than on development of aural/oral proficiency.
### Table 1. Postdictive Relationships for TU-J ESL Placement Variables with High School Grades in English as a Foreign Language and in all Academic Subjects, Respectively

<table>
<thead>
<tr>
<th>Year of assessment</th>
<th>1989-90</th>
<th>1990-91</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1 (N = 820)</td>
<td>(a)</td>
<td>(b-a)</td>
</tr>
<tr>
<td>Sample 2 (N = 828)</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Variable</td>
<td>Academic GPA</td>
<td>EFL GPA</td>
</tr>
<tr>
<td>H.S.GPA</td>
<td>.15</td>
<td>.27</td>
</tr>
<tr>
<td>SLEP LC</td>
<td>.24</td>
<td>.37</td>
</tr>
<tr>
<td>SLEP Total</td>
<td>.21</td>
<td>.36</td>
</tr>
<tr>
<td>Essay</td>
<td>.20</td>
<td>.35</td>
</tr>
<tr>
<td>Interview</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td>Placement composite*</td>
<td>.23</td>
<td>.39</td>
</tr>
</tbody>
</table>

Note. Findings in this table are from internal TU-J analyses conducted by Kimberly Graves for the present study sample.

* Weighted sum of z-scaled (M=0,SD=1) scores for SLEP Total, Essay and Interview: \(0.5 \times \text{Total} + 0.25 \times \text{Essay} + 0.25 \times \text{Interview}\).

### Objectives of the Present Study

The present study focused on more specific validity-related properties of the SLEP test by

(a) analyzing relationships of SLEP summary scores--and subscores reflecting performance on their component item types--with the two direct measures in light of theoretical expectation and empirical findings involving similar measures in other contexts, and

(b) conducting an exploratory analysis of intercorrelations among "item type parcels" (subsets of four to 10 or more items of the respective types), using factor analytic methods, to address questions regarding the extent to which items included in the two SLEP sections tend to exhibit internally consistent patterns of interrelationships--e.g., that common factors interpretable as corresponding to "listening comprehension" and "reading comprehension", respectively, will be identified by the corresponding SLEP item types.

A related, but incidental objective was to evaluate patterns of relative performance on item-type subtests for TU-J students in light of patterns exhibited by native-English speaking students (e.g., Holloway, 1984); also those of unpublished findings (e.g., ETS, 1980, 1988) for
ESL samples used in developing different forms of the SLEP--analyses designed to shed light on largely uninvestigated questions bearing on the "diagnostic potential" of different types of SLEP items;\(^2\)

**Organization of this Report**

Generally speaking, analytical procedures employed in each of the foregoing lines of inquiry are described in detail in connection with the presentation of related findings. A description of the study context and the study variables, immediately following, precedes an evaluation of sample performance on the primary study variables--that is, LC, RC, and Total scores on the SLEP, interview rating and essay rating--as well as on SLEP item-type subtests. Presented next are analyses concerned with the concurrent and discriminant validity properties of SLEP summary scores--and subscores reflecting performance on items of the respective types included in the SLEP--with respect to the two direct measures (interview and essay). The analysis of intercorrelations among SLEP item-type parcels is then considered. Factor analytic procedures and related findings are described in detail. The report then concludes with an evaluative review of study findings and several suggestions for further inquiry.

**Study Context and Data**

Temple University-Japan (TU-J) offers a four-year liberal arts program, attended primarily by native-Japanese speaking students, in which academic instruction is conducted in English. As a matter of national policy, English study is required of all Japanese students during the six middle- and high-school years--approximately three hours of academic instruction per week for 35 weeks per year. Average levels of proficiency in TU-J samples can be thought of as reflecting levels attained after some 800 classroom-hours of instruction in English--a common element in the language learning histories of the students involved.

Curriculum-based instruction in Japanese schools tends to give greater emphasis to the development of reading skills than to the development of oral English communication skills (e.g., Saegusa 1983, 1985). In these circumstances, TU-J students may tend to be more somewhat more advanced (i.e., more "native-like"), on the average, in reading than in listening comprehension.\(^3\)

---

\(^2\) In the report of findings of a survey of SLEP users (Wilson, 1993: pp. 39-40), attention was called to the interest of SLEP users in questions regarding the validity of SLEP item types for predicting basic performance criteria (e.g., ratings of oral language proficiency or writing ability), and questions regarding the diagnostic potential of subscores based on subsets of SLEP items—and the dearth of research bearing on such questions.

\(^3\) Such a pattern of differential development—that is, development of functional listening comprehension skills lagging behind that of, say, reading comprehension skills—was documented by Carroll (1967) for a national sample of college-senior level students majoring in a foreign language (French, Spanish, German and Russian, respectively) in the United States, and may tend to be common to academic, second (foreign) language acquisition contexts. For further development of this point see Wilson 1989 (pp. 11-18; 52-53).
In any event, academically qualified applicants for admission to TU-J must meet minimal ESL proficiency requirements, and all such applicants are tested for English proficiency using the battery of indirect and direct assessment procedures outlined below:

- the Secondary Level English Proficiency (SLEP) test,
- an interview for assessing oral English proficiency, typically conducted by master's level ESL instructors and evaluated using locally devised scoring procedures,
- a writing sample on topics developed locally, rated holistically according to guidelines adapted from those developed by ETS for scoring the Test of Written English (TWE)--see, for example, ETS (1993).

The SLEP and the writing test are administered in a morning session; interviews are conducted in the afternoon session. To place students in one of five IELP-defined proficiency levels for ESL instructional purposes, a weighted composite of the total SLEP score (sum of scaled LC and RC scores) and the two ratings is used--after z-scaled transformation (M=0,SD=1), SLEP Total score and scores on the two direct measures are combined with weights of .50, .25, and .25, for SLEP Total, interview rating and essay rating, respectively.

The data employed in this study were generated in 16 operational testing sessions conducted during 1989-90 and 1990-91, involving some 1,646 TU-J applicants. Eight testing sessions were conducted between November and August of each academic year (one each during the November to April period, and one each in July and August). The data included scaled section and total scores on the SLEP, item-level response data, and observations on the two direct measures for each member of the study sample.

Characteristics of the SLEP

The SLEP is a norm-referenced test containing a total of 150 multiple-choice questions in two sections of 75 items each (for a detailed description see, for example, ETS, 1991). The first section is designed to measure listening comprehension and the second is designed to measure reading comprehension. The time required for the entire test is approximately 85 minutes: just under 40 minutes, paced by recorded prompts, for the listening section and 45 minutes for the reading section. Three equated forms of the SLEP (Forms 1, 2, and 3) are currently available. Test booklets are reusable. Both self-scoring and machine-scorable answer sheets are available.

Number right raw scores for each section are converted to a common scale with a minimum value of 10 and a theoretically obtainable maximum value of 40. However, in practice, the range of converted scores is more limited, ranging from 10 to 32 on the LC section and from 10 to 35 for the RC section (e.g., ETS, 1991, p. 16). The total score is the simple sum of the two converted section scores. Thus, in practice, total converted scores can range between 20 and 67. Reliability coefficients above the .9 level are reported for the respective section scores and the total score (e.g., ETS, 1991).

The test is composed of eight different item types, as enumerated in Table 2. Appendix A provides descriptions and illustrative examples of items of each of the types indicated in Table 2, and a brief descriptive overview will be provided later when analyses involving these item types
are considered. In including different types of items within a measure the test developer ordinarily assumes that the different item types represent primarily different methods of measuring, or tend to tap somewhat different aspects of, the same general underlying functional ability, skill or proficiency.

The Direct Measures

The two direct measures are described briefly below (and more fully in Appendix A).

**Writing Sample.** Examinees are given 25 minutes to prepare an essay on a given topic. Dictionaries are not used. The instructions (read aloud) tell examinees to "try to write on every line for a full page." They are encouraged to spend about five minutes thinking about the topic and making notes if they wish before they begin writing. Essays are read by full-time ESL instructors, all of whom have master's degrees in ESL or a related field.

The readers read and score 20 or 40 of these exams, usually in one sitting. Each essay is read by two readers and scored holistically on a six-point scale--guidelines developed by ETS for scoring the Test of Written English or TWE (see, for example, ETS, 1993). If the two ratings differ by no more than one point, the final score is the simple sum of the two scores; somewhat more complex averaging rules are followed when scores differ by more than one point. The range of possible final scores is 2 (e.g., two ratings of "1") to 12 (e.g., two ratings of "6"), corresponding conceptually to Levels 1 through 6 as defined for scoring the TWE--and corresponding numerically to those levels when divided by two.

**Oral English Proficiency Interview.** Each examinee is interviewed by one rater, privately, for six to ten minutes. Interviewers/raters are full-time ESL staff, with no special training in the interview/rating procedure, who interview six or seven examinees per hour for two hours. A skills-matrix model is employed in evaluating the interviewee's performance. The skills involved are labeled communicative behavior, listening, mean length of utterance, vocabulary, grammar, and punctuation/fluency. Each skill is rated separately on a scale ranging from 1 to 7+, in which the respective intervals are defined by brief behavioral descriptions. The overall score is the simple average of ratings on the respective components; only the overall score is used for placement.

Sample Performance on Study Variables

Descriptive statistics for the primary study variables (SLEP LC, RC and Total scaled scores, interview rating, and essay rating) were computed for samples defined by SLEP form taken (Form 1, 2, or 3), and for the combined sample. SLEP performance patterns in the TU-J sample were evaluated relative to those for SLEP reference groups. Because TU-J staff rated essays in terms of the behaviorally defined levels used for rating the TWE it was possible to make tentative comparisons of the distribution of TWE-scaled ratings for TU-J students with distributions reported by Educational Testing Service (ETS) for examinees taking the TWE in connection with plans to study in the United States (e.g., ETS, 1992, 1993).
Table 2. Items of Designated Types in Current Forms of the SLEP

<table>
<thead>
<tr>
<th>Item type</th>
<th>Number of items</th>
<th>Form 1</th>
<th>Form 2</th>
<th>Form 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Picture</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Dictation</td>
<td>20</td>
<td>19</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Maps</td>
<td>12</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Conversations</td>
<td>18</td>
<td>20</td>
<td>20*</td>
<td></td>
</tr>
<tr>
<td>(LC )</td>
<td>(75)</td>
<td>(75)</td>
<td>(75)</td>
<td></td>
</tr>
<tr>
<td>Cartoons</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Four Pictures</td>
<td>15</td>
<td>16</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Cloze**</td>
<td>40</td>
<td>39</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Completion</td>
<td>(22)</td>
<td>(22)</td>
<td>(30)</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>(18)</td>
<td>(17)</td>
<td>(14)</td>
<td></td>
</tr>
<tr>
<td>Reading Passage</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>(RC)</td>
<td>(75)</td>
<td>(75)</td>
<td>(75)</td>
<td></td>
</tr>
</tbody>
</table>

Note. The number of items of the respective types differs slightly by form. In each form of the SLEP, the LC and RC item types appear in the order in which they are listed, above.

* In Form 1 and Form 2 all of the recorded conversations involve three people. In Form 3, the first 14 conversation items involve three people, but the last six items involve only two people.

** These items are not treated separately in ETS internal test analyses. "Completion" items simply require examinees to pick options that complete a sentence correctly and "Comprehension items require them to answer questions based on a completed passage. See Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988, pp. 10-12) for discussion of this type of distinction in the context of providing a more complex classification of multiple-choice cloze items similar in type to those used in the SLEP.

To permit evaluation of performance on the SLEP item-types, mean percent right scores were computed for the eight basic types of items enumerated in Table 2, above, namely, Single Picture, Dictation, Maps, Conversations, Cartoons, Four Pictures, Cloze and Reading Passages, respectively. The relative performance of TU-J students was evaluated in light of the performance of a sample of native-English speaking students (G7-12) tested by Holloway (1984), and other samples of ESL students (e.g., ETS, 1980, 1988; Butler, 1989).
In evaluating the performance of TU-J students it is useful to recall that during three middle-school years Japanese students typically take EFL courses three hours per week for 35 weeks, and during high school, over a three-year period, typically 5 hours per week for 35 weeks (see Wilson, 1989, Endnote 16 for detail). Thus, for the great majority of sample members the performance under consideration reflects outcomes associated with required "common exposure" to some 800 hours of curriculum-based instruction in English as a foreign language (EFL). Moreover, in the Japanese secondary-school EFL curriculum, development of reading and writing skills in English reportedly (e.g., Saegusa, 1985; Saegusa, 1983) receives more emphasis than does development of EFL oral/aural skills.

Attention is directed first to evaluation of TU-J performance on the basic placement battery, and then to evaluation of performance on SLEP item types.

Performance on the Placement Battery

Table 3 shows means and standard deviations for SLEP LC, RC and Total scaled scores, interview rating, and essay ratings respectively, in the three SLEP-form samples. As has long been recognized (e.g., Carroll, 1967) scores on a norm-referenced test do not convey directly information regarding the functional abilities of individuals who attain them, although some such information may be provided by percentile ranks, if the level of functional ability in the normative sample(s) is known.

Table 4 provides some interpretive perspective re level of ESL functional ability associated with various SLEP scores, in the form of percentile ranks in the basic SLEP reference group (for example, ETS, 1991), for students classified according to ESL placement levels in U.S. secondary schools involved. For example, the SLEP Total mean of 40, for TU-J students, suggests that their average level of proficiency is slightly above that for "full-time ESL" placement and somewhat lower than that for "part-time ESL" placement in U. S. high school ESL programs.

Also shown in Table 4 is the "expected TOEFL Total scaled score" corresponding to mean SLEP Total scores such as those shown in the table (as reported by ETS, 1997, Table 16, p. 27). Note that the TU-J sample SLEP Total mean (Mn=40) reportedly corresponds to a TOEFL Total mean of 375--below the 10th percentile for undergraduate-level TOEFL examinees, for

---

4 Questions regarding the general level of ESL proficiency acquired by students in the sample are of interest and appear to be relevant. For example, the developmental level at which a given sample is functioning in a target language may affect the level and patterning of relationships among measures of different aspects of proficiency (see, for example, Hale et al., 1988; Oltman, Stricker, and Barrows, 1988). Direct assessment of such effects is beyond the scope of the present study.

5 Results of multiple discriminant analyses, not shown, indicate statistically significant differences among the samples with respect to the set of variables under consideration, but the differences do not appear to reflect pragmatically important differences in level of proficiency.
Table 3. Descriptive Statistics for the TU-J Sample, by Form of SLEP Taken

<table>
<thead>
<tr>
<th>Variable</th>
<th>Form 1 (N=684)</th>
<th>Form 2 (N=619)</th>
<th>Form 3 (N=342)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
</tr>
<tr>
<td>Oral English</td>
<td>3.9</td>
<td>1.5</td>
<td>4.1</td>
</tr>
<tr>
<td>Writing*</td>
<td>4.3</td>
<td>1.8</td>
<td>4.6</td>
</tr>
<tr>
<td>(Scale equivalent)</td>
<td>(2.2)</td>
<td></td>
<td>(2.3)</td>
</tr>
<tr>
<td>SLEP LC</td>
<td>18.1</td>
<td>4.0</td>
<td>18.8</td>
</tr>
<tr>
<td>SLEP RC</td>
<td>21.1</td>
<td>3.5</td>
<td>21.4</td>
</tr>
<tr>
<td>SLEP Total</td>
<td>39.2</td>
<td>6.9</td>
<td>40.2</td>
</tr>
</tbody>
</table>

Note: Data are for students tested for placement in the TU-J Intensive English Language Program during 1989-90 and 1990-91.

* Each essay is rated on a scale involving only six levels (descriptions provided in Appendix B), but the "writing score" is a composite of two ratings--three when the two raters differed by more than one point and a third rating fell between the first two (the third rating only was accepted if one of the first two scores differed by two points from the other and the third rater's score). Scores vary between 2 (e.g., both readers assign Level 1 ratings) and 12 (e.g., both readers assign Level 6 ratings). Thus, by inference, the means reported here averaged approximately at Level 2 on the basic TWE-parallel rating schedule (as indicated by the values in parentheses).
Table 4. SLEP Total Score Performance for ESL Placement Classifications from the Basic SLEP Validation Study*

<table>
<thead>
<tr>
<th>Classification</th>
<th>SLEP Total score</th>
<th>Mean</th>
<th>S.D.</th>
<th>Percentile</th>
<th>Estimated TOEFL**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-time bilingual prog</td>
<td></td>
<td>32</td>
<td>9</td>
<td>22</td>
<td>(300)</td>
</tr>
<tr>
<td>Part-time bilingual prog</td>
<td></td>
<td>37</td>
<td>11</td>
<td>33</td>
<td>(350)</td>
</tr>
<tr>
<td>Full-time ESL program</td>
<td></td>
<td>38</td>
<td>10</td>
<td>36</td>
<td>(350)</td>
</tr>
<tr>
<td>(TU-Japan, all forms)</td>
<td></td>
<td>40</td>
<td>7</td>
<td>42</td>
<td>(375)</td>
</tr>
<tr>
<td>Part-time ESL program</td>
<td></td>
<td>43</td>
<td>11</td>
<td>49</td>
<td>(400)</td>
</tr>
<tr>
<td>&quot;Mainstream&quot;</td>
<td></td>
<td>50</td>
<td>12</td>
<td>66</td>
<td>(475)</td>
</tr>
</tbody>
</table>

* Classification for instructional purposes in secondary schools participating in the basic SLEP validation study (Stansfield, 1984; ETS, 1987; ETS, 1997, Table 4). Several SLEP-survey (Wilson, 1992) respondents independently reported that academically qualified ESL students who earn a SLEP Total score approximately equal to, or higher than 50, generally are ready (in terms of English proficiency) to participate full-time in an English-medium academic program at the secondary-school level.

** From the SLEP Test Manual (ETS, 1997: Table 16) which provides average "expected TOEFL Total scaled score" associated with selected values of SLEP Total scaled score. The estimates shown here involve extrapolations of the average TOEFL Total score, for SLEP mean values not specifically tabled.

Formal behavioral descriptions of linguistic proficiency corresponding to the "ESL placement" categories shown in the table were not available. However, the two TU-J scales provide formal behavioral descriptions, and accordingly permit inferences regarding the functional levels of ESL speaking proficiency and writing ability represented in the sample. In addition, because essay ratings were rendered according to the same scale as that used for rating performance on the Test of Written English (TWE), it is possible to make useful, albeit tentative, comparisons of the writing skills of TU-J students with those of international students planning to study in the United States who take the TOEFL and the TWE in connection with such plans. And, by using a general regression equation for predicting TWE score from TOEFL scores, adapted from DeMauro (1993), it was possible to make an indirect assessment of the extent to which TU-J's IELP staff adhered to the TWE scoring guidelines. This was done by comparing the mean observed (TWE-scaled) essay rating with the "predicted TWE mean" corresponding to a TOEFL Total score of 375--the ETS-estimated average for SLEP examinees with a SLEP Total score of 40, the average average attained by TU-J students (see Table 3 and related discussion, above).
Levels of Functioning: Speaking and Writing

To characterize behaviorally typical levels of "speaking proficiency" for TU-J students, a general composite description of the type of linguistic behavior associated with the mean TU-J interview rating was developed by combining descriptions for designated component-skill ratings by level—e.g., the Level 3 descriptions for Listening, Vocabulary, Fluency and so on. Table 5 provides such a summary (composite) description for ratings at TU-J Level 3 and Level 4, respectively—representing the two most frequently attained levels in the sample. Findings not shown in the table indicate that some 60 percent of the sample earned interview scores below 4.5; some 46 percent earned scores between 2.5 and 4.4 (corresponding to Levels 3 and 4, respectively, for which composite skills descriptions are shown in Table 5).

Table 6 shows descriptions of essays rated at Level 2 (modal for TU-J) and Level 3 (relatively high for TU-J). As noted above, ratings were rendered by TU-J staff according to level descriptions developed by ETS for scoring the TWE (see, for example, ETS, 1993). Accordingly it is of interest to compare distributions of TWE-scaled essay ratings for TU-J students with distributions of TWE-scaled essay ratings for with distributions of TWE ratings that are reported by ETS for TOEFL/TWE examinees, especially examinees at "lower" levels of TOEFL-assessed ESL proficiency. Level 4, respectively—representing the two most frequently attained levels in the sample.

Such a comparison is provided in Figure 1, which portrays the distribution of TWE-scaled ratings for TU-J students and the distribution of TWE scores for TOEFL/TWE examinees in each of three TOEFL Total score ranges, including those with TOEFL Total scores below 477—the lowest of several TOEFL-Total categories for which TWE-score distributions have been reported by Educational Testing Service (e.g., ETS, 1992).

Note in Figure 1 that TWE-scaled ratings for TU-J students (with an estimated TOEFL Total mean = 375) center approximately at Level 2, while those for TOEFL/TWE candidates who scored below 477 on the TOEFL (that is, at or below the 30th centile for TOEFL undergraduate candidates) center at about TWE Level 3. No TOEFL Total mean was reported for the latter group.

Generally speaking, the direction of the difference in level between the two distributions is consistent with expectation based on (a) the low estimated TOEFL Total mean of 375 for the TU-J sample (see Table 4 and related discussion), and (b) empirical findings indicating that TOEFL Total score and TWE score tend to be moderately strongly related in samples of TOEFL/TWE examinees (e.g., DeMauro, 1992; ETS, 1993).
Table 5. Behavioral Descriptions Corresponding to Interview Performance Rated at Levels 3 and 4 According to the TU-J Scale: Levels Attained Most Frequently by TU-J Students

<table>
<thead>
<tr>
<th>Skill</th>
<th>Description for interview at TUJ-Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening.</td>
<td>The student appears to understand simple questions in common topic areas, but does not respond to topic changes, (and) requires nearly constant repetition and restatement;</td>
</tr>
<tr>
<td>Grammar.</td>
<td>Grammar of short utterances is usually correct or early correct, (but there is) no attempt at any utterance other than simple present;</td>
</tr>
<tr>
<td>Vocabulary.</td>
<td>Limited vocabulary, but some knowledge of appropriate ative vocabulary in limited areas (e.g., family, sports);</td>
</tr>
<tr>
<td>Mean length of utterance.</td>
<td>Provides short sentences, even if not correctly formed.</td>
</tr>
<tr>
<td>Communicative behavior.</td>
<td>Responds to attempts by interviewer to communicate, but seems reluctant to respond;</td>
</tr>
<tr>
<td>Pronunciation/ fluency.</td>
<td>There is no continuous speech to evaluate (although) pronunciation of simple words and phrases is clear enough to understand.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Skill</th>
<th>Description for interview at TUJ-Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening.</td>
<td>The student can understand and infer interviewer's questions using simple verb tenses (present and past tense), and seems to understand enough vocabulary and grammar of spoken English to allow for basic conversation;</td>
</tr>
<tr>
<td>Grammar.</td>
<td>Can use simple present and simple past tense with some accuracy, though frequent mistakes can be noted;</td>
</tr>
<tr>
<td>Vocabulary.</td>
<td>Attempts to use what vocabulary he/she has although frequently inaccurate;</td>
</tr>
<tr>
<td>Mean length of utterance.</td>
<td>Can give one or two sentence responses to questions;</td>
</tr>
<tr>
<td>Communicative behavior.</td>
<td>Attempts to maintain contact and seeks clarification.</td>
</tr>
<tr>
<td>Pronunciation/ fluency.</td>
<td>Pronunciation and fluency problems cause frequent communication problems . . . but) there are rudiments of a speaking 'style'.</td>
</tr>
</tbody>
</table>
Table 6. Descriptions of Essays at TWE Level 2 and TWE Level 3

<table>
<thead>
<tr>
<th>Level 2.</th>
<th>&quot;Demonstrates some developing competence in writing, but remains flawed on both the rhetorical and syntactic level.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A paper in this category may reveal one or more of the following weaknesses:</td>
</tr>
<tr>
<td></td>
<td>- inadequate organization or development</td>
</tr>
<tr>
<td></td>
<td>- failure to support or illustrate generalizations with appropriate or sufficient detail</td>
</tr>
<tr>
<td></td>
<td>- an accumulation of errors in sentence structure and/or usage</td>
</tr>
<tr>
<td></td>
<td>- a noticeably inappropriate choice of words or word forms</td>
</tr>
<tr>
<td>Level 3.</td>
<td>Almost demonstrates minimal competence in writing, but is slightly flawed on either the rhetorical or syntactical level, or both.</td>
</tr>
<tr>
<td></td>
<td>A paper in this category</td>
</tr>
<tr>
<td></td>
<td>- shows some organization, but remains flawed in some way</td>
</tr>
<tr>
<td></td>
<td>- falls slightly short of addressing the topic adequately</td>
</tr>
<tr>
<td></td>
<td>- supports some parts of the essay with sufficient detail, but fails to supply enough detail in other parts</td>
</tr>
<tr>
<td></td>
<td>- may contain some serious errors that occasionally obscure meaning</td>
</tr>
<tr>
<td></td>
<td>- shows absolutely no sophistication in language or thought.&quot;</td>
</tr>
</tbody>
</table>
The foregoing considerations suggest general validity for the relatively low average placement of the TU-J sample with respect to "level of ESL writing ability" as indexed by essay ratings rendered by TU-J staff according to TWE scoring guidelines. However, they do not permit inferences regarding the probable degree of comparability between levels of writing ability inferable from the TU-J distribution of TWE-scaled ratings--of locally developed writing samples, evaluated locally according to TWE scoring guidelines--and levels inferable from distributions of TWE scores such as those portrayed in the figure. Evidence bearing somewhat more directly on this issue was obtained by using a regression equation (adapted from DeMauro, 1992) for predicting TWE score from TOEFL Total score, to determine the average TWE score that would be expected for U.S.-bound students from the Asian region who obtain TOEFL Total

DeMauro (1993) reported descriptive statistics, correlations, and regression results for scores on the TOEFL and scores on the TWE, in each of eight international TOEFL administrations, separately for each of three TOEFL-program defined regions. For present purposes, results for the Asian region were deemed to be most pertinent. Unweighted averages of eight Asian-administration TOEFL Total means and TWE means, respectively (DeMauro, 1993, Table 3) were computed, as were the corresponding regression slope and intercept values (ibid., Table 6). The resulting averages for TOEFL and TWE were, respectively, approximately 512 and 3.50; the average slope and intercept values are reflected in the following equation: $\text{TWE} = (0.0101875 \times \text{Total}) - 1.715$. This equation was used to estimate the TWE score corresponding to the estimated TOEFL Total score of 375 for the combined TU-J sample.
scores averaging 375. When the imputed TOEFL mean (375) for the TU-J sample was substituted in the equation, the resulting estimated TWE score (2.1) was very close to the mean observed TWE-scaled essay rating (2.3).

Such close agreement between these two means provides more specific, albeit still indirect, evidence suggesting that IELP-staff adhered closely to TWE scoring guidelines in rating the local writing samples.  

Performance on SLEP Item Types

As noted earlier (see Table 2, above, and related discussion), summary scores for listening comprehension and reading comprehension, respectively, reflect performance on several different types of items which are described briefly below (see Appendix A for illustrative items of each type).

Listening Comprehension Item Types

Single Picture items require students to match one of four recorded sentences with a picture in the test booklet; Dictation items call for matching a sentence printed in the test booklet with a sentence heard on tape; in completing Map items, examinees refer to a map printed in the test book which shows four cars variously positioned relative to variety of possible destinations, street names, and so on (e.g., "first avenue", "government building", "library"), in order to identify the one car that is the source of a brief recorded conversation (see Appendix A); Conversations items require examinees to answer questions after listening to conversations recorded by American high school students.

Reading Comprehension Item Types

Cartoons items require examinees to match the reaction of one of four characters in a cartoon with one of five printed sentences; Four Picture items require examinees to examine four drawings and identify the one drawing which is best described by a printed sentence; in Cloze Completion items, students must complete passages by selecting appropriate words or phrases from among four choices printed at intervals in the passages, while in Cloze Comprehension items, students must answer questions about the passage for which he or she supplied the missing words or phrases; Reading Passage items require students to answer several printed questions after reading a relatively short passage (10 to 15 lines).

Performance Findings

The psychometric properties of subscores by type of item are routinely evaluated and reported internally at ETS in samples assessed for purposes of developing different forms of a

---

7 Moreover, these findings attest—abeyt incidentally and indirectly—to (a) the validity of the SLEP/TOEFL-Total correspondencies reported by ETS (1997) and (b) the potential generalizability of the regression-based guidelines reported by DeMauro to samples of ESL users/learners, from the Asian region, who take the SLEP.
standardized test, such as the SLEP (e.g., ETS, 1980), but little attention ordinarily is given to further study of such subscores. Since only summary scores are equated across test forms, subsets of items by type are not constrained to be "parallel" either within or across test forms. And in the case of the SLEP, the number of items by type is not constant across test forms. However, use of percent-right scores for the respective item types provides a useful basis for inferences regarding relative ease or difficulty of the items involved, especially inferences regarding subgroup differences in patterns of relative difficulty, the principal focus of the present analysis.

Examination of mean percent right scores for the eight item types described above for the TU-J sample, indicated substantial consistency in the rank ordering of item types according to the percent-right index of relative difficulty in each of three form-samples--there was only one reversal in rank order (involving Cartoons and Four Pictures subtests in Form 3).

Given similarity in patterns across the three form-samples, attention can be focused on results for the combined sample--that is, mean percent-right scores for the respective item type subtests without regard to test form. These are shown in Table 7 for TU-J students, along with corresponding means for a sample of native-English speaking (E1) high-school students enrolled in several Florida (USA) schools (Holloway, 1984), and for designated ESL samples.

Profiles based on the means in Table 7 are shown in Figure 2. The lower horizontal line in the figure reflects "middle difficulty" (approximately 62 percent correct) for SLEP items--more generally, the midpoint between the maximum possible score on a set of items, and the score that would be expected if each item were answered at random for a 4-choice test such as the SLEP; the upper horizontal line marks the "90 percent correct" level for a subtest, a value selected arbitrarily to represent "mastery" of the skills measured by a given test in a given sample.

It can be seen in the figure that the native-English speaking sample performed at or near "mastery" level on seven of the eight SLEP item type subtests: 90 percent or more correct for LC items, regardless of type; also for Cartoons and Four Pictures items--and approximately 80 percent correct for Cloze items as well; Reading Passages items were of "average difficulty" for the native-English speaking sample.

Given the perspective provided by the native-speaker-performance profile, it is apparent that the sub-skills being tapped by SLEP item types are quite unevenly developed in the TU-J sample (as well as in the other ESL samples, in which the respective item types appear to exhibit about the same pattern of relative difficulty as that portrayed for the TU-J sample). More specifically, for example, we may infer that, on the average, the ability of TU-J students to comprehend connected, conversational discourse in English (e.g., as in Conversations or Maps items) is less well developed than is the ability to recognize the written counterpart of a sentence-level utterance (as in Dictation items) or the ability to comprehend sentence-level, context-embedded written matter (as in Cartoons and Four Picture items).

The Reading Passages subtest was especially difficult for TU-J students (who averaged only 28 percent correct). This subtest was of "middle difficulty" for the native-English speaking
Table 7. Performance of the TU-J Sample, Other E2 Samples and a Sample of Native-English Speakers, on SLEP Item Types

<table>
<thead>
<tr>
<th>Item-type subtest</th>
<th>Mean percent right for samples</th>
<th>Native-speakers*</th>
<th>TU-J</th>
<th>ETS**</th>
<th>LACCD***</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mn</td>
<td>Mn Rank</td>
<td>Mn Rank</td>
<td>Mn Rank</td>
</tr>
<tr>
<td>Single picture</td>
<td></td>
<td>97</td>
<td>61 (4)</td>
<td>83 (3)</td>
<td>68 (2.5)</td>
</tr>
<tr>
<td>Dictation</td>
<td></td>
<td>93</td>
<td>80 (2)</td>
<td>87 (2)</td>
<td>67 (4)</td>
</tr>
<tr>
<td>Map</td>
<td></td>
<td>97</td>
<td>41 (6)</td>
<td>72 (5)</td>
<td>60 (5)</td>
</tr>
<tr>
<td>Conversation</td>
<td></td>
<td>93</td>
<td>35 (7)</td>
<td>66 (6)</td>
<td>52 (6)</td>
</tr>
<tr>
<td>Cartoon</td>
<td></td>
<td>94</td>
<td>85 (1)</td>
<td>88 (1)</td>
<td>77 (1)</td>
</tr>
<tr>
<td>Four pictures</td>
<td></td>
<td>91</td>
<td>79 (3)</td>
<td>76 (4)</td>
<td>68 (2.5)</td>
</tr>
<tr>
<td>Cloze</td>
<td></td>
<td>81</td>
<td>46 (5)</td>
<td>62 (7)</td>
<td>49 (7)</td>
</tr>
<tr>
<td>Literary Passage</td>
<td></td>
<td>64</td>
<td>28 (8)</td>
<td>52 (8)</td>
<td>29 (8)</td>
</tr>
</tbody>
</table>

Note. In evaluating the low percent-right means for the literary passage items, consider the fact that the reading passage and its associated set of 8 items is found at the end of the timed reading section. Thus, scores on the reading passage items may reflect items-not-reached variance, which tends to reflect differences in speed of responding to test items.

* Native-speaker data are from Holloway (1984)--for native-English speaking students in several Florida high schools (G7-12).

** ETS data are from unpublished internal analyses for international-student (ESL) samples used in test development and equating (ETS, 1980, 1988). The mean shown is the unweighted average of means for the two linguistically heterogeneous samples involved, made of up students in the G7-12 grade range.

*** LACCD data are from Butler (1989), and reflect the average performance of students assessed for ESL placement by nine colleges that comprise the Los Angeles Community College District (LACCD). The data tabled for the ETS sample are unweighted averages of means reported by ETS for two samples of international students involved in the development of SLEP Form 1 and SLEP Form 2 (ETS, 1980; 1988), respectively; data tabled for LACCD are means for ESL students entering nine member colleges of the Los Angeles Community College District (LACCD), tested by Butler (1989). Following each of the subtest means for the three ESL samples is the within-sample rank position of the mean relative to that of other subtest means.
sample as well. In evaluating the difficulty level of this subtest, it is pertinent to note that the reading passage and its associated set of eight questions are found at the end of the timed reading section. Data not tabled indicate that almost 25 percent of the TU-J sample received a score of "0" on Reading Passages subtest. Thus, it is possible that the score on the reading passage items may reflect some items-not-reached (INR) variance (that is, individual differences in speed of verbal processing).

The foregoing findings—indicating differential development of subskills represented by the SLEP item types in the TU-J sample (and other ESL samples)—suggest that the corresponding subscores may have potential value for diagnosis.

**Concurrent and Discriminant Validity of SLEP Scores**

For the purpose of evaluating SLEP’s concurrent and discriminant validity properties, intercorrelations of the principal study variables (SLEP LC, RC and Total, interview rating and essay rating) were analyzed in each of the three SLEP form-samples and in the combined sample

---

8 Letters refer to correspondingly lettered endnotes.
In addition, several regression analyses were also conducted to permit a systematic assessment of patterns of relationships among the variables under consideration. More specifically, interview rating, essay rating and the unweighted sum of the two ratings, respectively, were regressed on LC and RC treated as a set of independent variables. Then, LC and RC, respectively, were treated as dependent variables, and regressed on interview rating and essay rating, treated as a set of independent variables. Correlation and regression outcomes in the respective form-samples were very similar to those in the combined sample. Accordingly attention is focussed primarily on results for the combined sample.

Table 8a shows the pertinent correlational findings for the combined, all-forms sample (N = 1,646); results for the three form-samples are also included to permit general assessment of similarities in patterns for the respective form-samples, but unless otherwise noted all subsequent references to findings are to those for the total sample. Generally speaking, the coefficients reported in table 8a indicate moderately strong, concurrent validity for the SLEP scores with respect to the direct measures (e.g., coefficients above .60 for SLEP Total with respect to both criterion variables). With respect to discriminant validity, the pattern of coefficients for SLEP LC with the direct measures (higher for interview than for essay) as well as those for SLEP RC (higher for essay than for interview) is consistent with general expectation for such measures. However, when the essay criterion is considered, the concurrent RC/essay coefficient is not different from the LC/essay coefficient, inconsistent with expectation of somewhat closer relationships between reading and writing skills than between listening and writing skills.

A more rigorous view of these relationships is provided by the total-sample regression results shown in Table 8b. The first three analyses (labeled Analyses 1 through 3 in the table) highlight the relative contribution SLEP LC and RC, when treated as independent variables in regressions involving the interview, the essay and the unweighted sum of the two ratings, respectively, as dependent variables. In the remaining analyses (see Analyses 4 and 5 in the table) SLEP LC and RC, respectively, were treated as dependent variables, and were regressed on interview and essay, treated as a set of independent variables. The relative contribution of the respective sets of independent variables to estimation of the designated criterion measures is indicated by the relative sizes of the corresponding beta (standard partial regression) weights shown in the table.

Results of Analysis 1 and Analysis 2 indicate clearly that when the interview is the criterion, the unique contribution of LC (beta = .584) is almost three times that of RC (beta = .181), whereas when the essay criterion is considered, the corresponding betas (.353 vs. 345) like the zero-order coefficients involved (.581 and .579, corresponding to rounded values in Table 8a, above) are almost identical. Results of Analysis 3, involving the composite (interview+essay) criterion (betas = .475 and .302 for LC and RC) tend to be consistent with correlational patterns shown in Table 8a which suggest that development of "productive ESL skills" (speaking and writing) tends to be indexed more by level of listening comprehension (as measured by SLEP LC) than by level of reading comprehension (as measured by SLEP RC). The patterns of beta weights for interview and essay in Analyses 4 and 5 indicate relatively greater weight for interview than for essay when LC is the criterion, and relatively greater weight for essay than for interview when RC is the criterion. This pattern is consistent with the pattern
Table 8a. Intercorrelations of SLEP Scaled Scores and the Two Direct Measures by SLEP Form and in the Total, All-Forms Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Essay</th>
<th>SLEP score</th>
<th>SLEP Form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LC</td>
<td>RC</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>Interview</td>
<td>.60</td>
<td>.63</td>
<td>.53</td>
</tr>
<tr>
<td>Essay</td>
<td>.58</td>
<td>.58</td>
<td>.64</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>.66*</td>
<td>.92**</td>
</tr>
<tr>
<td>RC</td>
<td></td>
<td></td>
<td>.90**</td>
</tr>
<tr>
<td>Interview</td>
<td>.58</td>
<td>.63</td>
<td>.50</td>
</tr>
<tr>
<td>Interview</td>
<td>.63</td>
<td>.65</td>
<td>.53</td>
</tr>
<tr>
<td>Interview</td>
<td>.60</td>
<td>.65</td>
<td>.52</td>
</tr>
<tr>
<td>Essay</td>
<td></td>
<td>.54</td>
<td>.59</td>
</tr>
<tr>
<td>Essay</td>
<td>.63</td>
<td>.60</td>
<td>.68</td>
</tr>
<tr>
<td>Essay</td>
<td>.61</td>
<td>.62</td>
<td>.65</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>.66*</td>
<td>.92**</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>.64*</td>
<td>.90**</td>
</tr>
<tr>
<td>LC</td>
<td></td>
<td>.74*</td>
<td>.95**</td>
</tr>
<tr>
<td>RC</td>
<td></td>
<td></td>
<td>.90**</td>
</tr>
<tr>
<td>RC</td>
<td></td>
<td></td>
<td>.91**</td>
</tr>
<tr>
<td>RC</td>
<td></td>
<td></td>
<td>.92**</td>
</tr>
</tbody>
</table>

Note. Sample size by SLEP form is as follows: Form 1 (N=684), Form 2 (N=619), Form 3 (N=343), total sample, all forms (N=1,646).

* In two SLEP test-development samples, the corresponding LC/RC coefficients were approximately .78. The TU-J sample is more homogeneous with respect to the abilities measured by the SLEP than the test-development samples involved.

** These coefficients reflect part-whole correlation, hence are spuriously high.
Table 8b. Regression of Interview and Essay, Respectively, on SLEP LC and RC and Corresponding Reverse Regressions: Selected Total-Sample Results (N = 1,646)

<table>
<thead>
<tr>
<th>Analysis/Variables in analysis/Dependent vs. other</th>
<th>Selected regression results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta*</td>
</tr>
<tr>
<td>Analysis 1 Interview vs. LC &amp; RC</td>
<td></td>
</tr>
<tr>
<td>LC</td>
<td>.514</td>
</tr>
<tr>
<td>RC</td>
<td>.181</td>
</tr>
<tr>
<td>Analysis 2 Essay vs. LC &amp; RC</td>
<td></td>
</tr>
<tr>
<td>LC</td>
<td>.353</td>
</tr>
<tr>
<td>RC</td>
<td>.345</td>
</tr>
<tr>
<td>Analysis 3 Interview + Essay (sum) vs. LC &amp; RC</td>
<td></td>
</tr>
<tr>
<td>LC</td>
<td>.475</td>
</tr>
<tr>
<td>RC</td>
<td>.302</td>
</tr>
<tr>
<td>Analysis 4 LC vs. Interview &amp; Essay</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>.446</td>
</tr>
<tr>
<td>Essay</td>
<td>.312</td>
</tr>
<tr>
<td>Analysis 5 RC vs. Interview &amp; Essay</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>.271</td>
</tr>
<tr>
<td>Essay</td>
<td>.415</td>
</tr>
</tbody>
</table>

Note: Analyses 1 through 3 highlight the relative weighting of SLEP LC and SLEP RC, treated as independent variables, in regressions involving the interview, the essay and the sum of the two, respectively, as dependent variables. Analyses 4 and 5 highlight the relative weighting of the interview and essay, treated as independent variables, in regressions involving LC and RC, respectively, as independent variables.

* The standardard partial regression coefficient.
of zero-order coefficients shown in Table 8a—that is, LC/interview > LC/essay; RC/essay > RC/ interview. As indicated in Table 8b, all outcomes are highly significant statistically.

Except for the stronger than expected contribution of listening comprehension relative to that of reading ability in predicting concurrent performance on a measure of writing ability (essay rating), levels and patterns of correlations involving SLEP LC and RC scores and the two direct measures (interview and essay) are generally consistent with theoretical expectation. They also tend to be consistent with empirical correlational findings that have been reported for generally corresponding sections of a closely related test—the TOEFL—with interview-assessed ESL speaking proficiency (e.g., Clark and Swinton, 1979) and score on the Test of Written English (e.g., ETS, 1993; DeMauro, 1992), respectively.

With regard to test/interview relationships, Clark and Swinton (1979: p. 43) reported findings similar to those found here for the SLEP after analyzing relationships between TOEFL section scores and interview rating in a linguistically heterogeneous sample of international students in ESL training sites in the United States. TOEFL/interview coefficients for the respective TOEFL scores were as follows: Listening Comprehension (r = .65), Structure and Written Expression (r = .58), Reading Comprehension and Vocabulary (r = .61) and Total (r = .68). Much the same pattern of test/interview correlations has been reported (e.g., Wilson, 1989; Wilson and Stupak, 1998) for samples of educated, adult ESL users/learners in Japan (and elsewhere) who were assessed with both the Test of English for International Communication [TOEIC] (e.g., Woodford, 1992; ETS, 1986) and the formal, Language Proficiency Interview (LPI) procedure. 9

With regard to the essay criterion, correlations reported in table 7a, above, for SLEP LC and RC scores with the TWE-scaled, TU-J writing test are a bit lower than, but are similar in pattern to, TOEFL/TWE correlations reported (e.g., ETS, 1992, 1993; DeMauro, 1992) for samples of international students from the Asian region (predominately native-speakers of Japanese, Chinese and Korean, respectively). Based on data for eight Asian-region test administrations with corresponding sample sizes varying between roughly 12,000 and 42,000, median zero-order TOEFL/TWE coefficients for TOEFL Listening Comprehension and Reading Comprehension and Vocabulary and Total, respectively, were .64, .61 and .64 (ETS, 1992, p. 12) versus SLEP/essay coefficients of .58, .58 and .64 for SLEP LC, RC and Total, respectively (rounded, from Table 7a, above). It is noteworthy that for all regions other than the Asian region, TOEFL/TWE correlations tended to be higher for scores on the two TOEFL non-listening sections, especially Structure and Written Expression, than for the Listening Comprehension

9 The TOEIC (e.g., ETS, 1991a) is used by international corporations in Japan and more than 25 other countries to assess the English language listening comprehension and reading skills of personnel in jobs requiring the use of English as a second language. Generally speaking, higher correlations between a measure of listening comprehension and performance in language proficiency interviews, with little or no increase in correlation when a reading comprehension score is added to the listening score (to form a total score), is consistent with the underlying functional linkage between development of the abilities measured by the LC test (e.g., to comprehend utterances in English) and development of the more complex abilities assessed in the interview (e.g., to comprehend and produce utterances in English). See Bachman and Palmer, 1993, for further evidence interpretable as indicating the distinctness of speaking and reading abilities.
score. Based on the findings that have been reviewed, the pattern of SLEP/essay relationships as well as the pattern of TOEFL/TWE relationships may tend to reflect effects associated with patterns of English language acquisition and use—especially of ESL writing skills—that may be characteristic of Asian-region samples of ESL learners/users, but not such samples from other world regions.10

**Relationships Involving SLEP Item-Type Subtests**

The patterns of relationships involving the SLEP LC and RC summary scores and the two direct measures, above, are generally consistent with expectation and thus tend to extend evidence of the discriminant validity properties of the two SLEP sections.

To evaluate corresponding patterns of relationships involving SLEP item-type subscores with the two direct measures, the corresponding zero-order coefficients were computed within each of the three form-samples. The item-type scores—but not the criterion measures—were z-scaled (M=0, SD=1) within each sample for analysis in the combined-forms sample (N=1,646). The pattern of subtest/criterion correlations within the three form-samples was very similar to that for the combined-sample. Accordingly only the combined-sample results for item type subtests and the two direct measures, shown in Table 9, are considered here; findings for SLEP summary scores (from Table 8a) are included for perspective. A "(+)" following the part-score label indicates that the pattern of correlations involving that item-type score with the interview and essay, respectively, is consistent with that observed for the corresponding section score; a "(-)") indicates the opposite.

It can be seen that except for the Dictation subscore, the item-type scores exhibited patterns of discriminant validity paralleling those of the respective section scores. In the sole "section-domain inconsistent" finding, the LC-Dictation subscore correlated somewhat more closely with the essay rating than with the interview rating. In evaluating this outcome, it is useful to recall that Dictation items require an examinee to process up to four written sentences in order to find the one sentence that corresponds exactly to a sentence-level spoken stimulus.

Thus, on balance, findings involving SLEP summary section scores (LC and RC) with interview and essay rating tend to be generally consistent not only with theoretical expectation but also with empirical findings involving similar measures in other testing contexts; and findings summarized in Table 9, suggest patterns of discriminant validity for item-type sub-scores that tend to parallel patterns observed for the summary scores to which they contribute—and the single exception to this (expected) parallelism involves a "listening comprehension" item type (Dictation) which appears to call for relatively extensive processing of written response options in order to respond to a spoken prompt.

---

10 In research undertaken as part of the development and validation of the TWE (e.g., Carlson, Bridgeman, Camp, and Waanders, 1985) scores on the nonlistening sections of the TOEFL correlated more highly with essay rating than did scores on the listening section. The sample of international students involved was linguistically heterogeneous and not restricted to students from the “Asian region”.

24
Table 9. Correlations of SLEP Section and Total Scores, and SLEP Item-Type Subscores, with Interview and Essay Rating, Respectively for the SLEP Item-Types (N=1,646)

<table>
<thead>
<tr>
<th>SLEP score</th>
<th>Interview</th>
<th>Essay</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>.64</td>
<td>64</td>
</tr>
<tr>
<td>Listening Comprehension</td>
<td>.63</td>
<td>.58</td>
</tr>
<tr>
<td>Single Picture (+)</td>
<td>.55</td>
<td>.50</td>
</tr>
<tr>
<td>Dictation (-)</td>
<td>.34</td>
<td>.37</td>
</tr>
<tr>
<td>Maps (+)</td>
<td>.48</td>
<td>.40</td>
</tr>
<tr>
<td>Conversations (+)</td>
<td>.47</td>
<td>.42</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>.53</td>
<td>.58</td>
</tr>
<tr>
<td>Cartoons (+)</td>
<td>.28</td>
<td>.30</td>
</tr>
<tr>
<td>Four Pictures (+)</td>
<td>.30</td>
<td>.36</td>
</tr>
<tr>
<td>Cloze (+)</td>
<td>.44</td>
<td>.51</td>
</tr>
<tr>
<td>Literary (+)</td>
<td>.24</td>
<td>.26</td>
</tr>
<tr>
<td>Passage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The (unequated) item-type subscores were standardized by test form (z-scaled to mean=0, SD=1) for this combined-forms analysis, but scales for the SLEP scaled scores and the two direct measures were not similarly transformed.

* "+" = coefficients are consistent with expectation;
"-" = coefficients are inconsistent with expectation.

Factor Analysis of SLEP's Internal Structure

Factor analytic methods were used to analyze intercorrelations among parcels of SLEP items by type—for the present study, the parcels were subsets of odd-numbered and even-numbered items of the respective types. Scores were computed for 18 parcels representing performance on odd-numbered and even-numbered items of each type; the number of items per parcel ranged between four and 15. For the factor analysis, a distinction was made between "cloze completion" items (labeled "clz") and cloze "comprehension" (cloze "reading" or "clzr") items. The 18 parcels were labeled as follows: Pico (Single Picture, odd items), Pice (Single Picture, even items), Mapo (Map, odd items), Mape (Map, even items), and so on for Dicto, Dicte, Convo, Conve, Carto, Carte, Pic4o, Pic4e, Clzo, Clze, Clzro, Clzre, Reado, Reade.

---

11 Cloze completion (Clz) questions require examinees to complete passages by selecting the word or phrase among four choices printed at intervals in passages that can be inserted most meaningfully to replace a word or phrase omitted at the respective intervals; cloze "reading comprehension" (ClzR) questions require the student to select, among four options, the one that appropriately answers questions about the passage for which he or she supplied the missing words or phrases.
Intercorrelations of scores on these parcels were computed for samples defined by SLEP form taken, and in the combined sample. For the combined-sample analysis scores on the item-type parcels were z-scaled (M=0,SD=1) within each of the three form-samples prior to being pooled for analysis--thus the correlation matrix involved reflected the within-form relationships of direct interest for present purposes.12

Analytical Rationale

The factor analysis was exploratory in nature. At the same time, based on lines of reasoning and empirical evidence outlined below, it was expected a priori that two or more factors would be needed to account for the correlations and that, of the corresponding factors, at least two would be marked, respectively, by item-types from the listening and reading sections.

More specifically, it was reasoned that nonchance performance on items involving spoken stimulus material requires an examinee to exercise developed ability to comprehend spoken utterances of the type involved; similarly, nonchance performance on items involving only written stimulus material indicates some level of developed ability to read such material with comprehension. Moreover, both the SLEP and the TOEFL--to which SLEP scores have been found to be relatively strongly related (e.g., ETS, 1991; see also Table 3 and related discussion, above)--include item types with the same distinguishing surface properties.13

In addition, although no previous studies of SLEP's factor structure appear to have been conducted, a substantial body of empirical evidence bearing indirectly on TOEFL's factor structure has accrued from TOEFL sponsored research--see Hale, Rock, and Jirele, 1989, for a critical review of a number of major TOEFL factor studies in the context of their own confirmatory factor study of TOEFL's dimensionality. These studies have consistently identified a "listening comprehension" dimension, defined by items, regardless of type, from the corresponding test section, and one or more "reading-related" dimensions, each defined by item types that involve only English-language written matter--factors defined variously by item types that do not involve listening (including those measuring vocabulary and reading comprehension, for example). Such a distinction has also been found to obtain in studies involving models other than factor analysis for assessing the dimensionality of the TOEFL (e.g., McKinley and Way, 1992; Oltman, Stricker and Barrows, 1988).

---

12 As previously noted, items included in different test forms are not designed to be parallel, hence are not necessarily comparable in either difficulty or validity. Accordingly, in analyzing item-level scores from different test forms to evaluate a test's "internal consistency" or factor structure it is useful to standardize the unequated subscores within the respective form-samples--for example, to express item-type scores as deviations from sample means in within-form-sample standard deviation units before pooling the data for analysis. The resulting intercorrelations thus reflect only the within-form relationships that are of primary interest (see Wilson & Powers, 1994 for elaboration of the rationale for this approach).

13 Although TOEFL and SLEP are not parallel with respect to item-type composition, both tests include "listening" and "nonlistening" item types; and TOEFL and SLEP scores have been found to be relatively highly intercorrelated (e.g., ETS, 1991), as noted in the review of related studies, above.
Thus, taking into account the general surface similarities between the two tests and evidence suggesting that they are relatively strongly related empirically, based on the TOEFL findings just reviewed it was reasoned that SLEP's factor structure also should tend to include, at least, factors corresponding to the two general dimensions that have consistently emerged in analyses of the TOEFL--that is, a factor marked by item types from the listening comprehension section and a factor marked by item types from the reading comprehension section.

There are, of course, no exact methods for answering questions as to number of factors that are needed to account for patterns of intercorrelations in a matrix, and as indicated, above, no previous factor studies involving the SLEP appear to have been conducted. However, certain rule-of-thumb criteria for dealing with such questions--based on experience and empirical assessments of alternative criteria--are widely recognized (e.g. Harman, 1976; Rummel, 1970; Gorsuch, 1974; Cureton & D'Agostino, 1983; Stevens, 1986). Prominent among such rules of thumb is the so-called "Kaiser criterion" (as cited, for example, by Harman, 1976: p. 185, with reference to Kaiser, 1960):

"... a practical basis for finding the number of common factors that are necessary, reliable, and meaningful for explanation of correlations ... among variables ... is that the number of common factors should be equal to the number of eigenvalues greater than one of the correlation matrix (with unity in the diagonal)."

Other investigators (e.g., Hakstan, Rogers, & Cattell, 1982) have reported that the validity of the Kaiser criterion (based on evaluations of alternative solutions) appears to be enhanced in relatively large samples (e.g., 350 or more cases) in which the ratio of number of variables to number of factors is at or below the .30 level, and mean communalities average approximately .60.

The Kaiser criterion was selected for preliminary guidance regarding the "number of factors" issue, and principal components analyses were replicated in each form-sample and in the combined-forms (combined) sample. Salient findings are summarized in Table 10. In each sample, five eigenvalues were either greater than or quite close to one (.99 and .97 for the Form 2 and Form 3 samples). With a resulting factors-to-variables ratio of less than .3--that is, 5/18--and mean estimated communalities (not shown in the table) near the .60 level (.58 in the combined-sample analysis, for example), the consistent principal components findings across three relatively large samples suggested that five components should be retained for further analyses within the respective form-samples and in the combined sample. For added interpretive perspective, however, it was decided also to evaluate reduced models involving two, three, and four factors, respectively. Both orthogonal (Varimax) and oblique (Oblimin) rotations were requested.

For all models except that involving only two factors, convergence occurred after relatively few iterations in each of the form-samples and in the combined-forms sample. When a two-factor solution was requested, however, Oblimin consistently failed to converge after 25 iterations, the two Varimax factors which did emerge in fewer than 25 iterations were ill-defined, and the two-factor model was not considered further.
Table 10. Results of Principal Components Analysis by SLEP Form and in the Total, Combined-Forms TU-J Sample

<table>
<thead>
<tr>
<th>Form</th>
<th>Eigen-vector (N=684)</th>
<th>Eigen-value</th>
<th>Pct var</th>
<th>Cum % var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.54</td>
<td>30.8</td>
<td>30.8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.48</td>
<td>8.2</td>
<td>39.0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.22</td>
<td>6.8</td>
<td>45.7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.09</td>
<td>6.0</td>
<td>51.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.03</td>
<td>5.7</td>
<td>57.5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.90</td>
<td>5.0</td>
<td>62.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Form</th>
<th>Eigen-vector (N=619)</th>
<th>Eigen-value</th>
<th>Pct var</th>
<th>Cum % var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.62</td>
<td>31.2</td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.39</td>
<td>7.7</td>
<td>39.0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.22</td>
<td>6.8</td>
<td>45.7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.05</td>
<td>5.8</td>
<td>51.5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.99</td>
<td>5.5</td>
<td>57.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.88</td>
<td>4.9</td>
<td>61.9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Form</th>
<th>Eigen-vector (N=343)</th>
<th>Eigen-value</th>
<th>Pct var</th>
<th>Cum % var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.65</td>
<td>36.9</td>
<td>36.9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.46</td>
<td>8.1</td>
<td>45.0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.09</td>
<td>6.1</td>
<td>51.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.05</td>
<td>5.8</td>
<td>56.9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.97</td>
<td>5.4</td>
<td>62.3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.75</td>
<td>4.2</td>
<td>66.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Combined forms*</th>
<th>Eigen-vector (N = 1,646)</th>
<th>Eigen-value</th>
<th>Pct var</th>
<th>Cum % var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.76</td>
<td>32.0</td>
<td>32.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.41</td>
<td>7.8</td>
<td>39.8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.15</td>
<td>6.4</td>
<td>46.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.05</td>
<td>5.8</td>
<td>52.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>5.5</td>
<td>57.5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.84</td>
<td>4.7</td>
<td>62.2</td>
<td></td>
</tr>
</tbody>
</table>

* In this analysis, parcels were z-scaled (M=0,SD=1) in the respective form-samples prior to pooling.
Summary of Factor Outcomes

Five-factor outcomes in the three form-samples were generally consistent with those for the combined-forms sample—that is, the first factor to emerge was clearly marked by uniformly strong loadings for three sets of item type parcels from the LC section (all of the Single Picture, Maps, and Conversations parcels), and the second was equally strongly identified by subsets of items of three types from the RC section (all of the Four Pictures, Cloze Completion, and Cloze Reading parcels). Each of the remaining three factors was item-type specific, being marked, respectively, by uniformly strong loadings for sets of Dictation, Cartoons, and Reading parcels. In the only exception to the foregoing pattern, in the sample composed of examinees who took Form 2, Cartoons and Four Pictures parcels coalesced to form a relatively clearly marked factor, leaving Cloze Completion and Cloze Reading parcels to mark a "reading" factor distinct from that marked solely by Reading Passages parcels—which consistently marked a separate factor in all analyses. The three- and four-factor solutions yielded factor structures that were not as consistent across replications as that involving five factors, and in one or more samples the reduced models were characterized by departure(s) from structural simplicity—e.g., split loadings for parcels of the same type.

Table 11 shows, for the combined sample only, sorted loadings of .30 or greater for item-type parcels on orthogonal factors for the respective models. A factor clearly interpretable as reflecting "listening comprehension" is marked by Maps, Single Pictures and Conversations parcels in all the models and a "reading ability" factor, marked by Four Pictures, Cloze Completion and Cloze Reading parcels, is common to both the five-factor model and the four-factor model.

It can be seen in the table that the four-factor model differs from the five-factor model only with respect to the emergence of a factor defined by Cartoons and Dictation parcels. This factor emerged only in the combined-sample analysis, and it is difficult to interpret in terms of a distinctive underlying function.

In the three-factor model, simplicity of structure is lost: Cartoons and Dictation join Four Pictures and Cloze Completion parcels, but the Cloze Reading parcels then split their loadings between the factor thus defined and the factor identified primarily by the parcels of Reading Passages items.

It thus appears that rotations involving fewer than the five factors suggested by the Kaiser criterion do not generate "reduced structures" that reflect consistent and readily interpretable correlational "affinities" for the three item types that identify "unique" factors in the five-factor model. Although the four-factor model shown in Table 11 for the combined sample appears to emergence of a factor defined by Cartoons and Dictation parcels. This factor emerged only in the combined-sample analysis, and it is difficult to interpret in terms of a distinctive underlying function.

In the three-factor model, simplicity of structure is lost: Cartoons and Dictation join Four Pictures and Cloze Completion parcels, but the Cloze Reading parcels then split their loadings
Table 11. Sorted Loadings of Parcels on Orthogonal Factors As a Function of the Number of Factors Requested: Loadings of .30 or Higher

<table>
<thead>
<tr>
<th>Parcel</th>
<th>Five factors</th>
<th>Four factors</th>
<th>Three factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1  F2  F3  F4  F5</td>
<td>Parcel</td>
<td>F1  F2  F3  F4</td>
</tr>
<tr>
<td>zMapo</td>
<td>.66</td>
<td>zMapo</td>
<td>.66</td>
</tr>
<tr>
<td>zPico</td>
<td>.66</td>
<td>zPico</td>
<td>.65</td>
</tr>
<tr>
<td>zConve</td>
<td>.65</td>
<td>zConve</td>
<td>.65</td>
</tr>
<tr>
<td>zMape</td>
<td>.65</td>
<td>zMape</td>
<td>.65</td>
</tr>
<tr>
<td>zPice</td>
<td>.62 .31</td>
<td>zPice</td>
<td>.62 .30</td>
</tr>
<tr>
<td>zConvo</td>
<td>.61</td>
<td>zConvo</td>
<td>.61</td>
</tr>
<tr>
<td>zClze</td>
<td>.68</td>
<td>zClze</td>
<td>.67</td>
</tr>
<tr>
<td>zClzro</td>
<td>.67</td>
<td>zClzro</td>
<td>.66</td>
</tr>
<tr>
<td>zClzo</td>
<td>.66</td>
<td>zClzo</td>
<td>.66</td>
</tr>
<tr>
<td>zClzre</td>
<td>.64</td>
<td>zClzre</td>
<td>.64</td>
</tr>
<tr>
<td>zPic4e</td>
<td>.57</td>
<td>zPic4e</td>
<td>.56</td>
</tr>
<tr>
<td>zPic4o</td>
<td>.52</td>
<td>zPic4o</td>
<td>.52 .31</td>
</tr>
<tr>
<td>zDicto</td>
<td>.84</td>
<td>zDicto*</td>
<td>.67</td>
</tr>
<tr>
<td>zDicte</td>
<td>.81</td>
<td>zDicte*</td>
<td>.66</td>
</tr>
<tr>
<td>zCarte</td>
<td>.82</td>
<td>zCarte*</td>
<td>.66</td>
</tr>
<tr>
<td>zCarto</td>
<td>.81</td>
<td>zCarto*</td>
<td>.65</td>
</tr>
<tr>
<td>zReado</td>
<td>.83</td>
<td>zReado</td>
<td>.79</td>
</tr>
<tr>
<td>zReade</td>
<td>.81</td>
<td>zReade</td>
<td>.77</td>
</tr>
</tbody>
</table>

Note: "e" endings denote even-number items; "o" denotes odd-numbered items. Item types are Single Picture (Pic), Dictation (Dict), Maps (Map), Conversations (Conv), Cartoons (Cart), Four Pictures (Pic4), Cloze Comprehension (Clz), Cloze Reading (Clzr), and Reading Passages (Read), respectively. The parcels involved were z-scaled within the respective SLEP form-samples prior to being pooled for this combined-forms analysis.

* This factor did not emerge in any of the three form-sample analyses.
between the factor thus defined and the factor identified primarily by the parcels of Reading Passages items.

It thus appears that rotations involving fewer than the five factors suggested by the Kaiser criterion do not generate "reduced structures" that reflect consistent and readily interpretable correlational "affinities" for the three item types that identify "unique" factors in the five-factor model. Although the four-factor model shown in Table 11 for the combined sample appears to be characterized by structural simplicity, results involving four factors were relatively inconsistent across samples--e.g., as indicated in the table, the factor defined relatively clearly by Cartoons and Dictation in the combined-sample analysis did not emerge in any of the corresponding form-sample analyses.

In any event, the two general factors in the five-factor model appear to be interpretable as corresponding to the two general underlying functions posited in developing the SLEP, namely, "listening comprehension" (marked consistently by three of the four item types making up the listening section), and "reading ability" (marked consistently by three of five item types from the reading section).

A general "listening comprehension versus reading comprehension" interpretation is strengthened by the fact that, as was reported earlier (see Table 9, above, and related discussion), Single Picture, Map and Conversation subscores correlated more closely with interview rating than with essay rating, whereas the opposite was true for each of the four RC item types, including Reading Passages and Cartoons which defined item-specific factors. Only the Dictation item type--which involves demands on listening not made by any RC item type, and exhibits reading-related patterns of external relationships--was ambiguously "placed" by results of both external and internal analyses.

To assess possible effects on concurrent and discriminant validity of "recombining" item types in such a way as to reflect combinations suggested by the factor outcomes, modified SLEP LC and RC scores were computed. The modified LC score was based solely on the three item types that consistently defined a "listening comprehension" factor; the modified RC score included not only the four current reading-section item types but also the Dictation items (with reading-related external correlations).

Tentative indications of such effects are provided in Table 12 which shows zero-order correlations with interview rating and essay rating, respectively, for subscores (unequated across test forms) reflecting the modifications just described; also correlations involving raw scores (similarly unequated across forms) for currently defined SLEP sections are included for perspective, as are those involving the scaled SLEP section scores (which are equated across test forms). Correlations are for the combined-forms sample.

It can be seen in Table 12 that Modified LC--a listening score which excludes Dictation items--correlates as highly with Interview as does the formal, raw LC score--which includes Dictation. It can also be seen that the raw RC score--which does not include Dictation--correlates about as highly with Essay as does the Modified RC score (which includes Dictation items). At the same time, the Modified RC score correlates somewhat more closely with interview rating than does the raw RC score, consistent with the inclusion of Dictation, which
Table 12. Correlation of Scores for SLEP Item Types Grouped According to Study-Identified Classifications (Modified LC and RC scores) versus Current SLEP Section Scores

<table>
<thead>
<tr>
<th>Score</th>
<th>(a) Interview r</th>
<th>(b) Essay r</th>
<th>Difference (b - a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified LC</td>
<td>.62</td>
<td>.54</td>
<td>-.08</td>
</tr>
<tr>
<td>Modified RC</td>
<td>.56</td>
<td>.59</td>
<td>+.03</td>
</tr>
<tr>
<td>Raw Total</td>
<td>.63</td>
<td>.63</td>
<td>.00</td>
</tr>
</tbody>
</table>

\(^a\) Sum of subtests: Single Picture, Map, Conversation (item types that define the listening comprehension, common factor--see Table 11 and related discussion).

\(^b\) Raw number-right SLEP Listening Comprehension score: four item types that currently make up the SLEP Listening Comprehension section.

\(^c\) Sum of subtests: five item types (Four Picture, Cartoon, Cloze and Reading Passage items, all current RC, plus Dictation items--with reading demands that appear to be at least equal to demands on listening comprehension. This subscore reflects performance on item types that define a common factor interpretable as "reading comprehension" as well as item types marking unique factors which nonetheless exhibit reading-related external validity patterns. These are the item types associated with Factors 2, 3, 4 and 5, respectively: see Table 11 and related discussion, above).

\(^d\) Raw number-right SLEP Reading score.

...has a listening component, in the former but not in the latter. Some validity-attenuating effects associated with lack of equating of raw section scores (regular and modified versions) across test forms also can be discerned--e.g., coefficients for the SLEP scaled scores tend to be slightly larger than those for the corresponding raw scores.

In essence, it appears that the recombining the item types as indicated for the modified LC and RC composite scores--that is, post hoc scoring of Dictation as a "reading" item rather than a "listening" item--has the potential to enhance discriminant validity for "modified listening comprehension," at the cost of somewhat diminished discriminant validity for the post hoc "modified reading" composite, with little change in overall concurrent validity.
REVIEW AND EVALUATION OF FINDINGS

This study was undertaken to extend evidence bearing on SLEP's validity as a measure of ESL listening comprehension and reading skills in college-level samples, using data provided by Temple University-Japan (TU-J) for native-Japanese speaking students—primarily recent graduates of Japanese secondary-schools, who were tested for placement in intensive ESL training courses offered by TU-J's International English Language Program (IELP). TU-J provided, for some 1,600 students, scores on each of the placement measures: SLEP (LC, RC, and Total), an interview rating and an essay rating; and also a file containing responses of students to each of the eight item types that make up the SLEP (four for listening and four for reading).

Results of internal analyses by IELP staff (see Table 1, above) provide evidence of validity for the SLEP and the two direct measures as measures of ESL proficiency in the TU-J context. For example:

(a) each of the ESL placement measures tends to correlate more highly with grade point average (GPA) variables reflecting secondary-school grades in English-as-a-foreign-language (EFL) courses, only (EFL GPA) than with grades in all academic courses (General GPA), respectively, and

(b) IELP's ESL placement decisions based on these measures are judged to have a strong degree of pragmatic validity—that is, for example, initial ESL placement typically is judged by IELP staff to have been "appropriate", based on observation of student performance in the ESL courses to which they were initially assigned based on the placement composite.

Concurrent and Discriminant Validity Findings

Such evidence, of course, attests generally to the validity of SLEP and the two direct measures, as measures of ESL proficiency in the college-level, TU-J context. The present study has focused on more specific aspects of SLEP's validity, namely, the concurrent and discriminant validity of SLEP section scores—and scores on their respective component item types—with respect to ESL speaking proficiency and writing ability, respectively, as indexed by ratings for interviews and writing samples.

Findings that have been presented indicate moderately strong levels of concurrent validity for summary SLEP scores with respect to the two direct measures—for example, SLEP Total score correlated at the same, moderately strong level ($r = .64$) with both interview and essay, and corresponding correlations involving the respective SLEP sections were also relatively strong, ranging between .58 and .63. With regard to discriminant validity, findings of correlational and regression analyses (see Tables 8a and 8b, above) involving LC and RC scores with interview and essay, respectively, were generally consistent with expectation for such measures.

A finding of comparable validity for LC and RC with respect to the essay criterion was inconsistent with the general expectation of somewhat closer relationships between reading and
writing than between listening and writing. However, findings from the TOEFL testing context, reviewed in detail above, indicate that the same "inconsistent" pattern may tend to be characteristic of correlations involving generally similar measures (scores for TOEFL Listening Comprehension and Reading Comprehension and Vocabulary, respectively, versus TWE score) in samples of TOEFL/TWE examinees from the Asian region (made up predominately of native-speakers of Japanese, Chinese and Korean, respectively).

Accordingly, the apparently "inconsistent" SLEP outcome was interpreted as suggesting the possibility of somewhat stronger linkage between development of ESL listening comprehension skills and development of corresponding writing skills in samples from the Asian region than in samples from other world regions, reflecting effects possibly associated with regional- and/or native-language-related differences in patterns of English language acquisition and use.

Scores for three of four item types making up the SLEP LC section exhibited "section-consistent" patterns of correlation with interview and essay, as did scores on each of the four item types included in the reading section—that is, for LC item-type scores (except that for Dictation), score/interview correlations were higher than score/essay correlations, and the opposite correlational pattern emerged, consistently, for the four reading-section item-type subscores. Dictation items require examinees to process up to four written sentence-level answer-options in order to identify the one option that corresponds precisely to a sentence-level, spoken prompt. Thus its somewhat higher correlation with essay than with interview suggests that it may tend to as much a measure of reading skills as it is of listening skills.

The patterns of external correlations involving SLEP section scores and their component item types with the two direct measures lend support to the working proposition advanced at the outset, namely, that scores on the SLEP are likely to exhibit useful concurrent and discriminant validity properties in samples of college-level ESL users/learners, as well as in corresponding samples from the population of secondary-level ESL users/learners originally targeted for assessment with the test.

**SLEP's Internal Consistency and Dimensionality**

To assess SLEP's internal consistency and dimensionality, an analysis was made of intercorrelations among scores on "item-type parcels" (subsets of several items of a given type) reflecting performance on each of the types of items included in the SLEP. Exploratory factor analyses were replicated in each of three SLEP-form samples and in the combined sample. Based on the results that have been described in detail (see Tables 10 and 11, above, and related discussion) a model involving five orthogonal factors—two general ability factors and three item-type specific factors—was judged to constitute a more adequate and interpretable representation of the intercorrelations than did models involving fewer factors.

The relatively complex factor results—e.g., with three item-type-specific factors—were nonetheless deemed to be interpretable as reflecting primarily a general distinction between aspects of ESL listening comprehension (measured by Single Picture, Map and Conversation items which consistently coalesced to form one common factor) versus aspects of ESL reading
comprehension (measured "unambiguously" by Cloze Comprehension, Cloze Reading and Four Pictures item types which defined a second common factor).

The remaining three factors were relatively clearly marked by odd-numbered and even-numbered parcels of items of only one type, namely, Dictation, Cartoons and Reading Passage items, respectively. The emergence of these three item-type specific factors simply suggests that the three item types involved--two from the reading section and one from the listening section, all exhibiting "reading-related" patterns of external correlations with the interview and essay criteria--constitute sources of test variance that, for reasons that are not readily apparent, is relatively independent of that generated by other SLEP items, regardless of section, in the sample(s) here under consideration. In connection with the foregoing, it is noteworthy that these three item types have also been found to contribute similarly independent test variance in samples employed in developing alternative forms of the SLEP, based on results of unpublished internal analyses conducted by ETS (e.g., Angell, Gallagher & Haynie, 1988).

In any event, generally speaking, the emergence of item-type specific factors may reflect "level of difficulty" effects, method-of-assessment effects associated with differences in format or type of content, and/or more basic construct-related effects (for example, speed of responding to test items), and so on. In the present context, as noted earlier, items of the three types marking unique factors were either relatively easy or quite difficult for the TU-J sample (see mean percent-right scores in Table 9, above).

The Reading Passage items are located at the end of the timed Reading Comprehension section, with the potential for generating significant "items-not-reached" or "speed-related" variance. Dictation items appear to call for at least equal exercise of sentence-level reading skill and sentence-level listening comprehension--not true for any of the other SLEP item type regardless of test section. Thus, the Dictation item type--unique to the SLEP, among ETS-based ESL proficiency tests--may tend to generate item-type-specific variance associated with its (apparently) strong integrative mix of demands on both listening and reading skill.

An exploratory assessment was made of the discriminant validity properties of a modified LC score--from which the Dictation items were excluded--and a modified RC score which included the Dictation items along with the other current RC item types. Based on correlations involving the modified scores with the interview and the essay, respectively, it appeared that recombining the item types as indicated for the modified LC and RC composite scores--that is, post hoc scoring of Dictation as a "reading" item rather than a "listening" item--may have the potential to enhance discriminant validity for a "modified listening comprehension" section, at the cost of somewhat diminished discriminant validity for the post hoc "modified reading" section, with little effect on overall predictive validity.

Thus, findings based on analyses of SLEP's internal consistency and dimensionality, as well as those bearing on the concurrent and discriminant validity properties of SLEP section scores, tend to confirm and/or extend findings of research reviewed at the outset. On balance, the findings collectively appear to warrant the conclusion that scaled scores for the two SLEP sections as currently organized provide information that permits valid inferences regarding individual and group differences with respect to two relatively closely related but
psychometrically distinguishable aspects of ESL proficiency, namely, the functional "ability to comprehend utterances in English" and the corresponding ability "to read with comprehension a variety of material written in English"—and that these properties are likely to hold for SLEP's use in college-level samples, as well as in precollege-level samples from SLEP's originally targeted population, namely, secondary-level ESL users/learners.

**Related Findings**

Exploratory analyses were conducted to evaluate the possibility that subscores based on SLEP item types might have diagnostic potential. Mean percent-right scores for Single Picture, Dictation, Maps, Conversations, Cartoons, Four Pictures, Cloze and Reading Passage items, respectively, were also computed. For interpretive perspective, the profile of means for TU-J students was compared with the corresponding profile for a sample of native-English speaking secondary-level (G7-12 range) students (Holloway, 1984). The native-English speakers demonstrated "mastery level" performance (90 percent or more correct) for six of the eight SLEP item types (all but Cloze and Dictation). Cloze items were relatively easy and Reading Passage items were of "middle difficulty" for the G7-12 sample of native-English speakers.

TU-J students, in contrast, exhibited sharply differing levels of performance on the item type subtests. They performed at a comparatively low level on items involving comprehension of connected discourse, whether spoken (as in Maps and Conversations items—41 and 35 percent right, respectively) or written (as in Cloze and Reading Passage items—46 and 28 percent right, respectively); and they performed relatively better (80 to 85 percent right), though still below native-speaker level, on items involving sentence-level, context-embedded, "cognitively undemanding" material (as in Dictation, Cartoons and Four Picture items).

It appears that information regarding relative performance on SLEP item type subtests may be of potential value for assessing relative development (e.g., toward native-speaker levels) of the corresponding aspects of proficiency in different samples of ESL users/learners. Such differences in average level of performance constitute a necessary, but not sufficient condition for inferring "diagnostic potential" for the subtests involved.

**Incidental Findings**

Although assessing SLEP's validity-related properties has been the primary focus of this study, it is important to recognize that study findings also suggest substantial validity for the interview and the essay procedures which, along with the SLEP, are used by Temple University-Japan for ESL screening and placement. The information provided by these strongly face-valid, direct testing procedures appears to permit valid inferences regarding individual differences in ESL speaking proficiency and writing ability, respectively.¹⁴

---

¹⁴ Direct assessment of reliability was not possible for the two direct measures. However, by inference from observed moderately high levels of correlation with SLEP scores (with reported reliability exceeding the .9 level (e.g., ETS, 1997) the procedures used by TU-J staff for eliciting and rating both interview behavior and writing samples appear to have a "pragmatically useful" degree of reliability.
From an overall interpretive perspective, it is noteworthy that in scoring their locally developed writing tests, TUJ-staff used guidelines developed by ETS to score the Test of Written Expression, or TWE (e.g., ETS, 1993). Moreover, it appears that they were able to adhere relatively closely to those guidelines, based on detailed findings reported above (see especially Table 4, Figure 1 and related discussion). The findings suggest that inferences about level of ESL "writing ability" based on TWE-scaled ratings, by TU-J staff, of locally developed "writing tests", may tend to be comparable in validity to inferences based on assessments of writing ability that are generated through formal TWE procedures. It thus seems clear that by using the TWE scoring guidelines, TU-J staff substantially extended the range of potentially useful interpretive inferences based on their locally developed data.

TU-J interview and interview-rating procedures, involving separate scoring for six component subskills by only one interviewer-rater, generated ratings that exhibited the same level of concurrent correlation with SLEP Total score as did the essay rating and related procedures, which involved separate scoring of each writing sample by at least two staff members. This suggests somewhat greater "efficiency of assessment" for the interview procedure, as compared with the "writing sample" procedure. Generally speaking, procedures for assessing writing ability appear to pose problems (of domain specification and sampling, for example) in samples of native-English speakers as well as samples of ESL users/learners.¹⁵

The component-skills-matrix, summative approach used in rating interview performance results in an overall rating that is not linked to any established global scale for rating speaking proficiency. Based on results for TWE-scaled essays, interpretive perspective for interview ratings might well be similarly enhanced if interview behavior were to be evaluated according to such an established scale. See Lowe and Stansfield (1988a, 1988b) for detailed descriptions of two widely recognized scales (and related procedures) for rating speaking proficiency; also other basic macroskills.

SUGGESTIONS FOR FURTHER INQUIRY

On balance, SLEP scores for listening comprehension and reading comprehension, respectively, are providing information that permits valid inferences regarding the two relatively closely related but psychometrically distinguishable aspects of ESL proficiency that they were designed to measure.

Study findings raise questions regarding the "proficiency domain classification" of the Dictation item type. Although this item type appears to contribute to SLEP's overall validity, the mix of demands on listening and reading skills represented in this type of item may tend to attenuate discriminant validity properties of the current LC section. It would be useful to explore effects that might be associated with increasing the load on listening comprehension (and

¹⁵ For further development of this general point, see Wilson (1989: pp. 60-61, p. 74); for a research review of problems involved in assessing ESL writing proficiency, see Carlson et al. (1985); for a comprehensive analysis of the numerous problems involved in the direct assessment of "writing ability" generally, see Breland (1983).
possibly short-term memory as well) in Dictation items (e.g., by using a brief conversational exchange as the prompt).

In any event, questions regarding the "proficiency-domain classification" and overall functioning of the Dictation item type might usefully be considered in developing future forms of the SLEP. Also pertinent for test development purposes are study findings indicating the presence of internally inconsistent correlational patterns not only for Dictation items, but also for Cartoon and Reading Passage items—patterns that appear also to characterize these item types in linguistically heterogeneous samples of ESL users/learners involved in ETS test analyses of SLEP items. SLEP items of these three types were either quite easy or quite difficult for TU-J students, and, as currently formatted they exhibited reading-related patterns of external correlations in the TU-J setting.

Findings regarding the Reading Passage item type suggest the importance of research designed to assess possible effects associated with the end-of-test placement of the Reading Passage and related questions—by evaluating comparative performance on a clearly speeded and clearly unspeeded subset of these items. More generally, it would appear to be important to evaluate comparatively, validity-related effects that might be associated with differences in "speed" of verbal processing in a second language.

The average performance of TU-J students on the SLEP and the direct measures reflect outcomes in a sample in which members share a common core of exposure to curriculum-based, required study of English as a foreign language from middle-school forward. However, the levels of ESL proficiency reflected by the average SLEP scores attained by TU-J samples, of course, are not necessarily representative of levels of proficiency typically attained by Japanese secondary-school graduates. Research designed to ascertain the typical levels of ESL proficiency in representative samples of Japanese secondary-school graduates would shed useful light on "outcomes of EFL instruction" in Japanese middle- and secondary-schools—and provide information that should be of value to all concerned with EFL instruction.

Research designed to assess relationships involving scores on the SLEP, the two direct measures, indices of secondary school performance, and so on, and (a) performance in ESL training courses offered by the International English Language Program, as well as (b) performance in TU-J's English-medium academic program would appear to be useful. Moreover, periodic reassessment of a sample of TU-J students from one or more entering cohorts—with the SLEP and both direct measures—would make it possible to assess gains in the four related aspects of proficiency being tapped by the TU-J placement battery under the conditions represented by the TU-J program.
REFERENCES


40


APPENDIX

Appendix A. Description and Illustrative Examples of SLEP Item Types (Adapted from ETS, 1996)

SLEP Section I. Listening Comprehension (LC)

The LC section measures the ability to understand spoken English, using four different item types. The material is paced by a recording and thus all students have an opportunity to attempt all the items.

*Single Pictures.* The student must match one of four recorded sentences with a picture in the test book. The sentences are spoken only once and are not printed in the test book. Thus, Single Picture items are relatively pure measures of ability to comprehend spoken English. The items are described additionally as "... dealing with correct recognition of minimal pair contrasts, juncture, stress, sound clusters, tense, voice, prepositions, and vocabulary (ETS, 1986, p. 7).

*Dictation.* These questions appear to approximate the type of dictation exercises used frequently in English language classes: the student must choose the one sentence, from four that are printed in the test book, that matches exactly the sentence heard on tape.

*Map.* The test book shows a map on which streets and buildings are designated by name, and four cars at different locations are designated by letters (A, B, C, and D, respectively). The examinee must choose the one car that is the focus of a brief conversation on the recording. Map items are said to assess a "variety of linguistic, cultural, and pragmatic concepts including directions, recognition of building names and associated vocabulary, distance, and time" (ETS, 1986, p. 10).

*Conversations.* Questions are based on recorded conversations involving American high school students, thought of as representing typical secondary school situations. The conversations take place in various parts of a school and deal with events that typically occur in each location; also extracurricular activities, academic subjects, school closings. Printed answer options are provided for recorded questions.

SLEP Section II. Reading Comprehension (RC)

This section of the test is 40 minutes long and is designed to measure ability to understand written English. Two of the four item types in this section use either pictures only, or a combination of pictured and related written matter, as stimulus material. Scores on this generally timed section can be affected directly by speed of responding to test items.

*Cartoons.* A cartoon is shown that depicts several characters in a generally familiar social/familial situation. The cartoon includes one or two brief written ("bubble") representations of thoughts or utterances of the participants. Each item takes the form of a brief written statement (two or three sentences long, involving not more than approximately 40-45
words, including articles and prepositions) based on the cartoon, followed by a related question and multiple choice options (four).

*Four Pictures.* Each item presents four pictures and a brief statement that is accurate for only one of the pictures. The student must identify the picture (identified by A, B, C, or D) for which the printed statement is accurate.

*Cloze.* Two types of multiple-choice cloze items are identifiable, namely, (a) completion (Clz) questions requiring examinees to complete passages by selecting the word or phrase among four choices printed at intervals in passages that can be inserted most meaningfully to replace a word or phrase omitted at the respective intervals; and (b) comprehension (ClzR) questions requiring the student to select among four options, the one that appropriately answers questions about the passage for which he or she supplied the missing words or phrases.

*Reading Passage.* The student must read a short literary passage and answer questions about it. Of the item types that make up the Reading Comprehension section, the Reading Passages items most closely resemble the type of item included in reading comprehension tests designed for native-English speakers.

SAMPLE ITEMS OF THE FOREGOING TYPES ARE PROVIDED BELOW.
Section 1

The first section of the SLEP test measures ability to understand spoken English and is about 40 minutes long. It is divided into four parts, with four different types of questions.

Part A

For the first type of question, the student must match one of four recorded sentences with a picture in the test book. The sentences are spoken only once and are not printed in the test book. This part contains items dealing with correct recognition of minimal-pair contrasts, juncture, stress, sound clusters, tense, prepositions, and vocabulary.

Sample Questions

Note: Pictures are for illustrative purposes only. Actual pictures in the test book are larger than the samples in this brochure.

1. **On tape:**
   Look at the picture marked number 1.

   **On tape:**
   (A) The boy is typing on the keyboard.
   (B) No one is looking at the computer.
   (C) One of the girls is pointing to the screen.
   (D) They’re putting a computer into a box.

2. **On tape:**
   Look at the picture marked number 2.

   **On tape:**
   (A) They are taking off their sandals.
   (B) They are putting hats in a box.
   (C) They are dancing in a circle.
   (D) They are standing against the wall.
3. **On tape:**
   Look at the picture marked number 3.

   **On tape:**
   (A) She’s looking out of the car window.
   (B) She’s driving on the highway.
   (C) She’s opening the car door.
   (D) She’s rolling up the window.

4. **On tape:**
   Look at the picture marked number 4.

   **On tape:**
   (A) The flowers are growing outside.
   (B) The table has been cleared off.
   (C) Someone has eaten all the cake.
   (D) The table is filled with cakes and pies.

**Part B**

These questions approximate the type of dictation exercises used frequently in English language classes: the student must match a sentence printed in the test book with a sentence heard on the tape. The questions focus on the relationship between structure and meaning.

Sample Questions

1. **On tape:**
   The taller plants keep growing all summer long.

   **In test book:**
   (A) By the summer those plants will be much taller.
   (B) Those tall plants should be cut back when it’s warm.
   (C) Once it’s cold the plants won’t grow any taller.
   (D) The taller plants keep growing all summer long.
2. **On tape:**
   Is it too warm to wear a coat?

   **In test book:**
   (A) Aren’t you going to bring your coat?
   (B) Why didn’t I bring my coat?
   (C) Doesn’t that coat look warm?
   (D) Is it too warm to wear a coat?

3. **On tape:**
   Jorge can’t come over because he has a piano lesson.

   **In test book:**
   (A) Jorge can’t come over because he has a piano lesson.
   (B) Jorge comes over less often since he started playing the piano.
   (C) Jorge played the piano when he came over.
   (D) Jorge’s piano lesson is over, but he’s still not coming over.

4. **On tape:**
   Many people had already left the beach before the storm.

   **In test book:**
   (A) Few people were on the beach during the storm.
   (B) The storm did a lot of damage along the beach.
   (C) Many people had already left the beach before the storm.
   (D) Most people did not think there would be a storm.

5. **On tape:**
   You should have allowed more time to finish that book.

   **In test book:**
   (A) You will have to read that book soon.
   (B) You ought to finish that book on time.
   (C) You should have allowed more time to finish that book.
   (D) You could ask someone if they saw that book.

**Part C**

The questions in this part are based on conversations between students or announcements made by teachers or administrators in a school. The questions are given before the talks begin, allowing students to direct their attention to listening for the correct answer. The questions and answers are printed in the test book. For each question, students must choose one of four answers.
Sample Questions

1. On tape:
   (Narrator) Listen for the answer to the following question. What did the girl think the homework assignment was for math class? Here is the conversation.
   (Boy) Did you figure out the answer to problem number ten in the math homework?
   (Girl) Number ten? I thought we were only supposed to do the first eight problems.
   (Boy) That’s what the teacher said at the beginning of class, but right before the bell rang she changed the assignment to the first twelve problems.
   (Narrator) What did the girl think the homework assignment was for math class?

In the test book:

What did the girl think the homework assignment was for math class?

   (A) Only problem number 10
   (B) Problem numbers 1 through 8
   (C) Problem numbers 1 through 10
   (D) Problem numbers 1 through 12

2. On tape:
   (Narrator) Listen for the answer to the following question. Which bus will not be running this afternoon? Here is the announcement.
   (Man) Please excuse this interruption. There has been a change in the school bus routes this afternoon. The number five bus has a flat tire, so all students who normally take the number five bus will take the number two bus today. All those students who are changing to the number five bus must report to the office sometime during periods four, five, or six to sign a change-of-bus-route form so that the bus driver will know how many students will be on the bus.
   (Narrator) Which bus will not be running this afternoon?

In the test book:

Which bus will not be running this afternoon?

   (A) Number 2
   (B) Number 3
   (C) Number 5
   (D) Number 6
3. **On tape:**

   (Narrator) Listen for the answer to the following question. Where is Linda’s new job? Here is the conversation.
   
   (Boy) Hey Linda, I hear you got a job at the music store in the mall. Sounds like fun.
   
   (Girl) Actually, I was offered a job at the music store. I accepted a job as a teller at the bank.
   
   (Boy) Well, that makes sense. You’re planning to study finance in college, right?
   
   (Girl) Exactly. I thought it would be better to work at a place now that would prepare me for what I want to do in the future.
   
   (Narrator) Where is Linda’s new job?

   **In the test book:**

   Where is Linda’s new job?
   
   (A) At a music store  
   (B) At a restaurant  
   (C) At a college  
   (D) At a bank

---

**Part D**

The questions in this part are based on conversations recorded by American high school students that represent typical secondary school situations. The conversations take place in various parts of a school and deal with events that typically occur in each location. For each recorded question, the student must choose one of four answers printed in the test book.

Sample Questions

1. **On tape:**

   (Sam) Hi, Luisa.
   
   (Lusia) Oh, Sam. Hi. I missed you this morning in history class. Where were you?
   
   (Sam) That’s what I wanted to talk to you about. You see, my homeroom teacher had to pack up a bunch of science materials from our unit on electricity, so she asked me stick around and help her.
   
   (Narrator) Why did Sam miss history class?

   **In the test book:**

   (A) He was sick.
   (B) He missed the bus.
   (C) He was helping a teacher.
   (D) He was doing a science experiment.
On tape:

(Luisa) Did Mr. Jackson say it was okay to miss his class?
(Sam) Yeah—my homeroom teacher called him and asked if he minded. He said it was okay with him if it was okay with me.
(Luisa) Well, all we did was take notes about the Industrial Revolution.
(Sam) I know. That’s why I was looking for you. I wanted to see if I could look over your notes tonight. I can bring them back tomorrow.

In the test book:

Why was Sam looking for Luisa?

(A) To give her a note.
(B) To return her science book.
(C) To ask to borrow her notes.
(D) To see if Mr. Jackson was upset.

2. On tape:

(Luisa) Oh, sure. That's not a problem. I have my notebook right here in my ... uh oh.
(Sam) What’s wrong?
(Luisa) Well, I thought my notebook was in my backpack, but I must have left it in Mr. Jackson’s room. I’d better go get it.
(Sam) I’ll go with you.

In the test book:

Where are Sam and Luisa going?

(A) To the lunchroom.
(B) To the history teacher’s room.
(C) To the library.
(D) To Luisa’s locker.
Section 2

The second section of the test measures ability to understand written English and is 40 minutes long. The questions cover grammar, vocabulary, and reading comprehension. There are four parts to Section 2.

Part A

For each question in this part, the student must match the reaction of one of four characters in a cartoon with a printed sentence.

Sample Questions

1. The beach is a perfect place for a relaxing vacation.
2. I’ll finally have a chance to finish painting the garage.
3. I’m looking forward to catching up on my reading.
4. I hope we stay at a campground where I can go horseback riding.
5. I can’t wait to try out that new ride at the amusement park.
Part B

For the questions in this part, the student must match a printed sentence with one of four drawings. The particular focus of this item type is the use of prepositions, pronouns, adverbs, and numbers.

Sample Questions

1. One bird is sitting in a tree but two aren’t.

   ![A] ![B] ![C] ![D]

2. The bigger circle is in the lower left corner.

   ![A] ![B] ![C] ![D]

3. She is reaching up to get a box off the shelf.

   ![A] ![B] ![C] ![D]

4. They headed for shelter when it started to rain.

   ![A] ![B] ![C] ![D]
Part C

This part of Section 2 contains questions of two types. In one, the student must complete passages by selecting the appropriate words or phrases from a set of four choices printed at intervals in the passage.

Sample Passage and Questions

1. Most animals in the desert do not take in water from open sources like lakes or rivers. Some of these animals may never even come upon any type of open

   (A) so
   (B) despite
   (C) why
   (D) which

2. water, they have adopted other ways to obtain it. Some desert-dwelling

   (A) so
   (B) despite
   (C) why
   (D) which

3. toads, for example, can take in from dew-soaked soil directly through

   (A) its
   (B) their
   (C) his
   (D) one’s

4. skin. Another way some desert animals deal with a lack of water is to

   (A) as much as
   (B) as many as
   (C) too much
   (D) so much

5. drink possible when they find water in large quantities. For

   (A) to survive
   (B) have survived
   (C) surviving
   (D) survive

6. instance, after drinking all they can hold, some camels can for more than two weeks in the desert without any more water

   (A) to survive
   (B) have survived
   (C) surviving
   (D) survive
Sample Questions

7. What is the main topic of the passage?
   (A) Plants that live in the desert
   (B) Types of desert camels
   (C) How to make freshwater
   (D) Surviving with little water

8. The word “adopted” in line 2 is closest in meaning to
   (A) attracted
   (B) developed
   (C) lost
   (D) asked about

9. What helps camels survive in the desert?
   (A) They drink a lot of water at one time when they find it.
   (B) They rest for two weeks before trips in the desert.
   (C) They travel early in the morning, when they can soak up water from dew.
   (D) The people who travel with camels pack large amounts of water.

Part D

In this part of Section 2, the student must read a short literary passage and answer questions about it.

Sample Passage and Questions

We stopped to buy gas and to stretch our legs. We had left home early that morning and driven for what seemed like years. Now it was noon and the sun overhead was oppressive. The baby was crying. I wondered if we would ever reach our grandmother’s house. Father bought us bottles of something cool to drink. As we sat sipping our drinks beneath a shady tree, he began to tell us a story.

1. What time did the family stop?
   (A) In the morning
   (B) At noon
   (C) In the afternoon
   (D) At night
2. Where was the family going?
   (A) To the gas station
   (B) To visit their grandmother
   (C) Home from the store
   (D) Back from the park

3. What was the weather like?
   (A) Cold
   (B) Rainy
   (C) Windy
   (D) Hot

4. Why does the writer mention “years” (line 2)?
   (A) The family has been driving for a long time.
   (B) The father tells his story slowly.
   (C) The grandmother is very old.
   (D) The writer does not remember what happened.

5. When did the father tell the story?
   (A) As he was walking
   (B) After he bought the drinks
   (C) Before the family sat down
   (D) When the family returned to the car

### Answer Key for Sample Questions on Pages XX-XX

<table>
<thead>
<tr>
<th>Section 1</th>
<th>Part A</th>
<th>Part B</th>
<th>Part C</th>
<th>Part D</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. D</td>
<td>4. C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 2</th>
<th>Part A</th>
<th>Part B</th>
<th>Part C</th>
<th>Part D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>7. D</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8. B</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9. A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B. TU-J Direct Assessment Procedures

TU-J: Instructions to Interviewers/Raters
TO: Interviewers
FROM:
DATE:
RE: IELP Placement Test Interviews

The interview part of the placement test is an important part of the entire entrance procedure for TUJ's IELP. Interviews take place after the written test has been completed and are designed to find out the applicant's approximate speaking level in a conversation/interview situation. Their speaking ability does not necessarily correlate with other English skills, but is simply one more *measure of a student's overall proficiency.

PROCEDURE -- For your interview session, you will receive the following:

- procedure instructions
- Applicant Record Form with attached picture for each person you are interviewing
- blank matrix scales for recording students speaking ability
- complete list of interviewees, with your group designated
- interview scale

The interview usually takes 6-10 minutes. Needless to say, the students are very nervous about this, for it may be their first experience speaking face-to-face with a native speaker. As a result, the interviewer should try and get the student to relax by making small talk and moving into the suggested questions in as natural a way as possible. If you ask questions that the student has anticipated, you may suddenly get seemingly fluent "canned" answers, which will be easy to recognize.

When you finish the interview use the Interview Scale to determine the speaking ability in each particular section. Mark each score on the individual matrix scale and average. Write the average on the individual Applicant Record Form in Part IV (Interview) in the score blank. Next, it is important to note comments which you have about the student. For example, you should note if a person refuses to speak or has difficulties speaking perhaps because of shyness difficulties and not language difficulties. Also, note if the applicant has been abroad, as well as the length of the stay. Please note if the person exhibits social, psychological or physical problems. Etc., Etc. space on the Applicant Record Form

Our interpretation of the student's ability will, of necessity, be somewhat intuitive. Familiarize yourself with the descriptions for each level on the scale, and use your judgement. Next, it is important to sign your initials in the appropriate to place them with a score.

Try to make this experience as non-threatening as possible. Feel free to think up your own questions but take care in the manner in which you ask them. The complexity of the grammar is something to consider. Some suggested questions might be as follows:
Where do you live?
Tell me about one of your teachers.
What is the name of your high school? Tell me about it.
Tell me a little about your family.
Do you have any hobbies? Tell me about them.
What do you do when you have free time?
What are your favorite sports / singers / foods / writers / songs?
Have you ever been abroad? If so, tell me about your experience.
Why do you want to study in an American university?
If you had a lot of money or a lot of time, what would you do?
Tell me about someone you respect and why you respect them.
Do you have any specific plans for the future?
Do you prefer Tokyo or your hometown? Why?
What kind of person would be the ideal husband/wife?

Some Do's and Don'ts for interviewers:

Until you've made a judgement as to their level, don't prompt answers, either in Japanese or in English. It's important to give the student an opportunity to ask for repetition or clarification.

In the comments section, explain how the person presents himself; whether they're eager to talk, reticent, etc.

Don't correct or repeat responses.

Don't overuse phatic signals (uh-huh, I see, umm, etc.)

Allow for periods of silence, for students to formulate answers or to decide that they want to ask for clarification.

Rate the student's level of self-confidence.

Don't talk too much until you've already rated the student.

Don't take over the conversation or the topic.
INTERVIEW SCALE/ Draft Rost, 3/87

Matrix Categories

Communicative behavior

1 No attempt to communicate
2 No attempt to initiate conversation, little eye contact
3 Responds to attempts by interviewer to communicate, but seems reluctant to respond
4 Attempts to maintain contact and seeks clarification
5 Maintains contact throughout interview, seeks clarification, asks for repetition consistently
6 Assumes an active role in the conversation, attempts to communicate, expand discussion, even in limited English
7 Appears highly motivated, behaves naturally, responds appropriately, acts as "equal partner" in the discussion

Listening

1 Does not understand anything that is said or asked
2 Appears to understand a few high frequency vocabulary items and very short yes/no questions
3 Appears to understand simple questions in common topic areas, but does not respond to topic changes. Requires nearly constant repetition and restatement.
4 Can understand and infer interviewer's questions using simple verb tenses (present and past tense); seems to understand enough vocabulary and grammar of spoken English to allow for basic conversation.
5 Shows confidence in understanding interviewer's questions and statements and seeks clarification of areas of misunderstanding.
6 Can recognize own listening problems; has strategies for allowing conversation to continue in spite of minor problems of understanding.
7 Has ample receptive competence in grammar and vocabulary to allow a normal conversation to take place. Only infrequent lapses in understanding.
8 Appears to understand almost everything that interviewer asks or comments about.

Mean length of utterance

1 Does not speak, expect for occasional "yes" or "no"
2 Only provides one or two word responses (Yes ...I can...) often with broken grammar (Yes...I English)
3 Provides short sentences, even if not correctly formed
4 Can give one or two sentence responses to questions
5 Can provide basic narration consisting of three or more sentences
6 Shows confidence in speaking continuously, even if in broken English
7 Can speak continuously if necessary.
**Vocabulary**

1. Produces only Japanese or Japanese/English cognates
2. Extremely limited vocabulary prevents conversation from developing at all
3. Limited vocabulary, but some knowledge of appropriate vocabulary in limited areas (e.g. family, sports)
4. Attempts to use what vocabulary he/she has although frequently inaccurate.
5. Vocabulary is adequate to hold basic conversation in general areas
6. Good working vocabulary. Only occasional lack of approximate word or expression.
7. Vocabulary is adequate to hold a conversation on most topics

**Grammar**

1. No grammar
2. Telegraphic speech; subject-verb or verb-object-order only
3. Grammar of short utterances is usually correct or nearly correct; no attempt at any utterance other than simple present
4. Can use simple present and simple past tense with some accuracy, although frequent mistakes can be noted
5. Some use of perfect tenses, although frequently incorrect, and adequate use of simple tenses
6. Grammar is not a problem for understanding intended meaning,
7. Can use most grammatical patterns needed for conversation on general topics; only occasional mistakes which interfere with communication although range of grammatical control is still limited

**Pronunciation/Fluency**

1. Can't understand what interviewee is saying
2. Pronunciation problems make conversation stop frequently
3. Pronunciation of simple words and phrases is clear enough to understand; no continuous speech to evaluate
4. Pronunciation and fluency problems cause frequent communication problems, but there are rudiments of a speaking "style"
5. Basic command of English pronunciation, but falters with rhythm and intonation in continuous speech
6. Shows basic fluency, can maintain rhythm and conversational intonation
7. Confident in using spoken English, pronunciation and intonation seem natural
## EVALUATION MATRIX

### INTERVIEW EVALUATION MATRIX

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7+</th>
</tr>
</thead>
</table>

Communicative behavior
Listening
Mean length of utterance
Vocabulary
Grammar
Pronunciation/Fluency

Interviewer's Initials: _______  
Student's Number: _______  
Average (to nearest tenth)

---

## EVALUATION MATRIX

### INTERVIEW EVALUATION MATRIX

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7+</th>
</tr>
</thead>
</table>

Communicative behavior
Listening
Mean length of utterance
Vocabulary
Grammar
Pronunciation/Fluency

Interviewer's Initials: _______  
Student's Number: _______  
Average (to nearest tenth)

---

## EVALUATION MATRIX

### INTERVIEW EVALUATION MATRIX

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7+</th>
</tr>
</thead>
</table>

Communicative behavior
Listening
Mean length of utterance
Vocabulary
Grammar
Pronunciation/Fluency

Interviewer's Initials: _______  
Student's Number: _______  
Average (to nearest tenth)
ESSAY EXAM

Think and write about the following topic:

Why is it important to have friends?

You have 25 minutes for the English essay. Use 5 minutes to think about this topic. You may write notes on this paper. Then use 20 minutes and write an essay in English on the second paper. Try to write one full page. Write on each line. Your notes will not be graded, but your English essay will be graded. Please remember to write in English.

Instructions are repeated here in Japanese.

Notes:
PROGRAM-WIDE WRITING TEST--SCORING GUIDELINES *

Score

6  Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.

A paper in this category

-is well organized and well developed
-effectively addresses the writing task
-uses appropriate details to support a thesis or illustrate ideas.
-shows unity, coherence, and progression
-displays consistent facility in the use of language
-demonstrates syntactic variety and appropriate-word.choice
-shows more sophistication in language and thought than does a 5 paper

5  Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will have occasional errors.

A paper in this category

-is generally well organized and well developed, though it may have fewer details than does a 6 paper
-may address some parts of the task more effectively than others
-shows unity, coherence, and progression
-demonstrates some syntactic variety and range of vocabulary
-displays facility in language, though it may have more errors than does a 6 paper

4  Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.

A paper in this category,

-is adequately organized
-addresses the writing topic adequately but may slight parts of the task
-uses some details to support a thesis or illustrate ideas
-demonstrates adequate but undistinguished or inconsistent facility with syntax and usage
-may contain some serious errors that occasionally obscure meaning
-shows little or no sophistication in language or thought

3  Almost demonstrates minimal competencies in writing, but is slightly flawed on either the rhetorical or syntactic level, or both.

A paper in this category
-shows some organization, but, remains flawed in some way
-falls slightly short of addressing the topic adequately
-supports some parts of the essay with sufficient detail, but fails to supply enough
detail in other parts
-may contain some serious errors that occasionally obscure meaning
-shows absolutely no sophistication in language or thought

2 Demonstrates some developing competence in writing, but remains flawed on both
the rhetorical and syntactic level.

A paper in this category may exhibit one or more of the following weaknesses:

- inadequate organization or development
- failure to support or illustrate generalizations with appropriate or sufficient detail
- an accumulation of errors in sentence structure and/or usage
- a noticeably inappropriate choice of words or word forms

1 Suggests incompetence in writing.

A paper in this category is seriously flawed by one or more of the following weaknesses:

- failure to organize or develop
- little or no detail, or irrelevant specifics
- serious and frequent errors in usage or sentence structure
- serious problems with focus

A paper in this category may be illogical or incoherent, or may reveal the writer's
inability to comprehend the question. A paper that is severely underdeveloped also
falls into this category.

* This scoring guide is adapted from the TWE Scoring Guide and reprinted by permission of
Educational Testing Service, the copyright owner.
a. It is noteworthy that similar findings regarding native-speaker performance on "reading items" have been reported for the TOEFL (e.g., Angoff and Sharon, 1971; Johnson, 1977; Clark, 1977). It does not follow from the fact that the SLEP Reading Passage items are of moderate difficulty for native-English speakers, that these items are not measuring pragmatically important aspects of "English proficiency." It has been argued, for example, that "if native speakers can't do it, then it's not a language test, it's something else" (observation by Oller, in the discussion of Oller and Streiff, 1975, p. 88); similarly, if the aim is language testing (rather than intelligence testing), this means ". . . excluding items which are failed by natives whose education and status is equivalent to the test population (Ingram, E., 1968, p. 93). As indicated above, TOEFL reading comprehension items also have been found to be relatively difficult for native-English speaking college freshmen (e.g., Angoff and Sharon, 1971) and high-school seniors (e.g., Clark, 1977). In evaluating this phenomenon, Clark (1977, pp. 30-31) noted that if all test items that are relatively difficult for native speakers were to be eliminated (from the TOEFL), the result would be the exclusion of items measuring "a number of aspects of basic English structure that are at least subjectively indispensable for effective academic work at the undergraduate level." This appears to hold for the SLEP as well.

b. Although omitted responses to items located at the end of a timed test do not necessarily reflect "speed-related effects" (see, for example, Swinton and Powers (1976), various "items not reached" indices have long been used at ETS for test analysis purposes. In any event, some degree of "speededness" is not necessarily a "validity reducing" phenomenon--see arguments (e.g., Oller [1983: p. 105]; Harris, 1968) for the position that speed or fluency in the use of language constitutes an important element of language proficiency in a very general sense; see also Bernhardt, 1984: p. 47, for a similar point of view. Thus, it does not follow from the fact that the SLEP Reading Passage items--very difficult for TU-J students, perhaps due in part to the plausible possibility that scores reflect general speed of processing written material in English as well as "reading comprehension", per se--that these items are not measuring pragmatically important aspects of ESL proficiency. On the other hand, other scholars have argued, for example, that "if native speakers can't do it, then it's not a language test, it's something else" (observation by Oller, in the discussion of Oller and Streiff, 1975, p. 88); similarly, according to Ingram (1968, p. 93), if the aim is language testing (rather than intelligence testing), this means ". . . excluding items which are failed by natives whose education and status is equivalent to the test population."

c. In the study cited, after correction for attenuation due to unreliability, within-section coefficients reached the .90 level--that would suggest measurement of the same underlying functional ability--only for Single Picture, Maps and Conversations (LC subscores) and for Cloze and Four Pictures (RC subscores); corresponding coefficients for these subscores with all others, both within- and across-sections, were substantially below the .90 level.

Three of the eight types (Dicatation, Cartoons and Reading Passages) did not appear to be strongly identified with any of the other types:

for Dictation items, within-section and between-section corrected coefficients ranged
between .47 and .69;

corrected coefficients involving the Cartoons item type were uniformly low, both within- and across-sections, ranging between .27 and .47 with LC item types and between .44 and .53 with other reading item types;

for the Reading Passages items the highest corrected coefficient was with Cloze items (.80); other within-section coefficients involving this item type (.44 and .58) were either lower or no higher than the corresponding across-section coefficients (.58 [Dictation] to .72 [Conversation]).

In the test analysis sample, these three subtests were either considerably easier (Dictation, Cartoons) or relatively more difficult (Reading Passage) than the other item-type subtests, with correspondingly skewed score distributions--true as well for the present sample. Items-not-reached analyses suggested some degree of speededness for the timed reading comprehension section which would tend to affect primarily performance on the Reading Passages items--by inference from their location at the end of the timed reading section.

d. Widely recognized scales include the ILR Oral Proficiency Interview scale adopted by the Interagency Language Roundtable (ILR), and the modification of that scale developed jointly by Educational Testing Service and the Association of College Teachers of Foreign Languages (ACTFL) as a uniform basis for assessing foreign-language speaking proficiency in academic contexts (for general descriptions and historical perspective on development and use of these scales see Lowe and Stansfield, 1988a, 1988b). The ILR scale includes behaviorally described levels ranging from "Level 0" (no proficiency) through "Level 5" (proficiency equivalent to that of a native-English speaker). Because relatively few foreign language students exhibit proficiency much above Level 3 on the ILR scale, a modified version of that scale was developed collaboratively by ACTFL and ETS to permit finer discrimination at the lower end of the ILR scale--by adding several levels at the lower end of the ILR scale--and eliminating level-distinctions beyond ILR Level 3, defining all behavior at or above that described for ILR Level 3 as "Superior." See Carroll (1983) for consideration of some of the factors involved in measuring second-language attainment at higher levels of proficiency.

e. Generally speaking, systematic documentation of gains in proficiency based on instruction--a logically relevant goal in any second language learning context--does not appear to have been a high-priority goal for those concerned with second/foreign language learning/teaching/assessment in academic settings. Indeed, the most systematic and comprehensive assessments of foreign-language attainments following intensive instruction appear to be those conducted (e.g., Lett & O'Mara, 1990) within a military training context in the United States, under the aegis of the Defense Language Institute Foreign Language Center (DLIFLC). DLIFLC studies, such as that cited above, have documented end-of-training levels of proficiency associated with intensive programs of instruction in over 40 different languages being studied as foreign languages by native-English speaking members of the U.S. Armed Forces. That documentation includes evidence of marked differences, by target language, with respect to levels of proficiency attained in basic target-language macroskills, for native-English speaking military personnel, after comparable periods of intensive training--in analyses that introduced controls for both native-language verbal ability and "aptitude for foreign language learning", as well as other variables. No comparably comprehensive
set of data bearing on "end-of-training" status for nonnative-English speakers engaged in studying (and using) English as a foreign/second language appears to be available--in any event, none could be located in a variety of published and unpublished sources which were examined in connection with the preparation of this report. See Carroll (1967) for a landmark study of the foreign language attainments of college-senior-level foreign language majors in the United States.