DOCUMENT RESUME

ED 280 295 FL 016 560

AUTHOR Arth, Thomas O.
TITLE Revision of the Basic and Intermediate English Language Tests.
INSTITUTION Air Force Human Resources Lab., Brooks AFB, Tex. Manpower and Personnel Div.
PUB DATE Dec 86
NOTE 23p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Armed Forces; Comparative Analysis; Correlation; Difficulty Level; *English (Second Language); Foreign Countries; *Foreign Nationals; Government Employees; Language Proficiency; *Language Tests; Multiple Choice Tests; Reading Tests; *Test Construction; *Test Items; Test Reliability; Test Validity; Writing Skills

ABSTRACT

The process of revising and validating two English language tests used by the United States armed forces in hiring foreign nationals overseas is described. Development of the item banks and classification of items are outlined, and field testing in the United States and overseas is described. The tests were the basic and intermediate level language skills measures. The basic-level test's format and content were retained, but the number of items in each subtest was increased. The intermediate-level test was revised to include multiple-choice subtests for reading, writing, and listening and an oral interview. Field testing consisted of administration of both the original and the revised versions and comparison of results. High correlations were found between the revised tests and the original tests and other measures of validity. Use of the revised versions in place of the originals is recommended. (MSE)

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

AFHRL-TR-86-42

# AIR FORCE HUMAN RESOURCES LABORATORY

ED280295

# REVISION OF THE BASIC AND INTERMEDIATE ENGLISH LANGUAGE TESTS

Thomas O. Arth, 1Lt, USAF

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

December 1986

Interim Report for Period September 1983 - March 1986

Approved for public release; distribution is unlimited

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

U.S. Gov't

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

FL016560



BEST COPY AVAILABLE

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Scientific Advisor
Manpower and Personnel Division

DENNIS W. JARVI, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>AFHRL-TR-86-42 | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Manpower and Personnel Division<br>6b. OFFICE SYMBOL (*If applicable*)<br>AFHRL/MOAO | 7a. NAME OF MONITORING ORGANIZATION |
| 6c. ADDRESS (*City, State, and ZIP Code*)<br>Air Force Human Resources Laboratory<br>Brooks Air Force Base, Texas 78235-5601 | 7b ADDRESS (*City, State, and ZIP Code*) |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Air Force Human Resources Laboratory<br>8b. OFFICE SYMBOL (*If applicable*)<br>HQ AFHRL | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |
| 8c. ADDRESS (*City, State, and ZIP Code*)<br>Brooks Air Force Base, Texas 78235-5601 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7719; TASK NO. 18; WORK UNIT ACCESSION NO. 47 |

11. TITLE (*Include Security Classification*)
Revision of the Basic and Intermediate English Language Tests

12. PERSONAL AUTHOR(S)
Arth, Thomas O.

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (*Year, Month, Day*) | 15 PAGE COUNT |
|---|---|---|---|
| Interim | FROM Sep 83 TO Mar 86 | December 1986 | 22 |

16. SUPPLEMENTARY NOTATION
Revision of the Basic and Intermediate English Language Tests. This work was accomplished under TS Study Numbers 8499, 8518, 8605, 8638, 8911, and 9147.

| 17. COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB-GROUP |
| 05 | 09 | |
| | | |

18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
civilian tests; proficiency tests; writing; English language tests; reading; listening; selection tests

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

The purpose of this effort was to update the Basic and Intermediate English Language Tests (ELTs). These tests are used in hiring foreign nationals at overseas bases. Currently, the Basic ELT consists of reading, listening, writing, and speaking tests; the Intermediate ELT is composed of three parts: two sentence completion and one sentence matching.

The format for the revised Basic ELT remained unchanged, whereas the Intermediate ELT was revised to include multiple-choice tests for reading, writing, and listening, as well as a speaking test interview. The revised Basic and Intermediate ELTs were administered, along with the current tests, to basic trainees to determine whether knowledge of English alone was sufficient to answer these items. Then all ELTs were pretested on a sample of foreign students at the Defense Language Institute. The item pools were reduced for field testing of the Intermediate ELT and final item selection made for the Basic ELT. Field testing of the Basic ELT occurred at Howard AFB, Panama. Field testing for the Intermediate ELT occurred at 16 bases overseas.

The results showed that the revised ELTs correlate highly with the current ELTs and other measures of validity. It is recommended that the revised ELTs replace the ELTs currently in use.

| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT | 21. ABSTRACT SECURITY CLASSIFICATION | |
|---|---|---|
| ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS | Unclassified | |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Nancy A. Perrigo, Chief, STINFO Office | 22b. TELEPHONE (*Include Area Code*)<br>(512) 536-3877 | 22c. OFFICE SYMBOL<br>AFHRL/TSR |

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

## SUMMARY

The Basic and Intermediate English Language Tests (ELTs) are used to make decisions on hiring of foreign nationals at bases overseas. Occasionally, the tests are also used to determine bonus pay. The Basic and Intermediate ELTs were last revised in 1965 and 1967, respectively. Therefore, the Air Force Civilian Personnel Center requested that the Air Force Human Resources Laboratory update both ELTs.

The existing Basic ELT consisted of four tests: Writing, Listening, Reading, and Speaking. The existing Intermediate ELT consisted of three parts that essentially measured reading ability. Revision of these tests entailed making two forms each of the Basic and Intermediate ELTs, each of which would contain four tests (Writing, Listening, Reading, and Speaking) of 25 items each.

This project was accomplished in three phases. Phase 1 administered a replacement item pool to native English-speaking subjects. Air Force basic trainees were used in this phase; 348 were used for the Basic ELT item pool and 635 were given the Intermediate ELT item pool. Results showed the basic trainees missed very few items, demonstrating that knowledge of English alone was sufficient in answering these items. In Phase 2, these item pools were pretested on samples of 99 Defense Language Institute foreign students. The item pools were administered along with the existing tests to ensure a comparable level of difficulty between the old and new instruments. Final item selection for the Basic ELT and selection of items for the Intermediate ELT field test item pool were based on the results of this phase. The last phase involved field testing the items on foreign nationals presently working at bases overseas. These results confirmed that the revised Basic ELT could discriminate among lower English-language ability subjects and provided the basis for final item selection for the Intermediate ELT.

This project culminated in two forms each of the Basic and Intermediate ELTs. Each form contains four 25-point tests. The Basic and Intermediate ELTs were revised to be complementary instruments, each containing a Writing, Listening, Reading, and Speaking test, with the Intermediate ELT being more difficult than the Basic ELT.

In future research, it is recommended that these tests be normed on job applicants. These norms might then be used to decide whether to administer the Basic ELT or to administer the Intermediate ELT.

## PREFACE

This work was completed under Task 771918, Selection and Classification Technologies. which is part of a larger effort in Force Acquisition and Distribution. It was subsumed under work unit number 77191847. Development and Validation of Selection Methodologies. This work was begun in response to Request for Personnel Research (RPR) 83-05, Revision of Basic English Language Tests for Use in Overseas Testing.

I would like to express my appreciation to the illustrator of the pictures in the Basic English Language Test, Al Young. I would also like to thank George Vliet for reproducing the audio cassettes in the Basic and Intermediate Listening tests. Also, thanks are due the personnel in the Technical Services Division who conducted several sets of analyses for this project. These personnel include Jim Brazel, Tom Sackett, Harry Loveland, Bill Glasscock. and Rodger Shutt.

TABLE OF CONTENTS



LIST OF TABLES

# REVISION OF THE BASIC AND INTERMEDIATE ENGLISH LANGUAGE TESTS

## I. INTRODUCTION

The English Language Tests (ELTs) are used to test foreign nationals seeking employment at bases overseas on their English-language proficiency. There are currently two versions of ELTs: the Basic and the Intermediate.

The Basic ELT consists of a Speaking test and a Listening test (which are preceded by a Speaking and Listening Warm-Up Exercise) and a Reading test and a Writing test (preceded by a Reading and Writing Warm-Up Exercise). Testing times for each are as follows: Listening test - 3.3 minutes, Speaking Test - 5 minutes, Reading Test - 3 minutes, and Writing Test - 5 minutes. Answers and distractors in the multiple-choice Listening test are presented in picture form. The stems in the other three tests are given in picture form. Each test has 20 items. All four tests have two parallel forms.

The Intermediate ELT has three sections. Part I measures vocabulary and contains 30 items. Part II measures grammar and is made up of 27 fill-in-the-blank items. Part III has 23 items that measure reading comprehension. All items are of the multiple-choice type. Testing times for each section are as follows: Part I - 15 minutes, Part II - 15 minutes, and Part III - 20 minutes. As in the case of the Basic ELT, there are two parallel forms of the Intermediate ELT. Table 1 gives a description of the Basic and Intermediate ELTs.

Table 1. Construction of the Existing Basic and Intermediate ELTs

| Test | Stem characteristics | Response characteristics |
|---|---|---|
| Basic Listening | Spoken Sentence | Four-picture, multiple-choice |
| Basic Speaking | Picture | Free response |
| Basic Reading | Picture | Four-word, multiple-choice |
| Basic Writing | Picture | Supply missing word |
| Intermediate Part I | Underlined word in sentence<br>Word Analogy | Four-word, multiple-choice |
| Intermediate Part II | Missing word in sentence | Three-word, multiple-choice |
| Intermediate Part III | Whole sentence-Sentence Analogy | Four-sentence, multiple-choice |

Although the ELTs proved to be an effective screening device, several problems have become apparent. The last revision of the ELTs was made in 1967, and the currency of the tests is questionable. Also, due to the length of time the tests have been in the field, the issue of compromise has been raised. Finally, there is a lack of documented validation of the ELTs.

An attempt was made in this effort not only to update the ELTs but also to improve them. They were improved by measuring all facets of language ability. The use of a language has four components: listening, reading, speaking, and writing. Of the 10 tests of English-language proficiency described in Buros (1978), none appropriately tested all four components in adults,

although a number of studies have sought to suggest ways to improve language ability measurement. Hisama (1977b) defended the use of multiple measures in order to avoid mismeasurement in testing English as a second language. In order to increase the effectiveness of a test that measures reading and listening, Pike (197g) developed criterion measures of speaking and writing ability to supplement the test. Lombardo (1981) developed an assessment battery that measured receptive language (reading and listening). She concluded receptive area tests were valid measures of language proficiency since they were interrelated with expressive (writing and speaking) areas. The study went on to note, however, that the receptive area precedes the expressive area in the acquisition of language. From this finding, it seems receptive area tests are valid only with elementary-level examinees.

"Banding" has been proposed as an effective method of determining the level of English-language proficiency. This is a system where the level of proficiency is divided into bands, ranging from beginner to native speaker. Corbett (1980) stated that banding is most useful when the specific purpose for which the language is to be used can be specified. Good banding standards can be maintained by designing a variety of tests. This method is similar to the ELTs in that there are both elementary (Basic) and advanced (Intermediate) levels of the test.

The CLOZE procedure has been extensively researched and has been found to be a reliable, valid, and practical measure of English-language proficiency. This is a technique developed by Taylor (1953) where every nth word is deleted from a paragraph. The examinee then supplies the missing word. Stubbs and Tucker (1974) validated a CLOZE test with an English proficiency entrance examination with excellent results. The CLOZE procedure was compared to several measures of English-language proficiency by Hisama (1977a) and was found to be both reliable and valid.

CLOZE tests have also been used in a multiple-choice format. Scholz and Scholz (1981) found open-ended and multiple-choice CLOZE tests appeared similar in their relationship to general English proficiency. Although multiple-choice tests have been criticized, they are a viable means of testing language proficiency. Schulz (1977) determined that objective, multiple-choice tests were more useful than simulated conversation tests as instructional aids for learning a foreign language.

Speaking tests are the most difficult to administer and score of all the language proficiency tests. This is due to the fact that they are somewhat subjective in nature. Subjectivity can be reduced by using the average of two judges' ratings, according to Mullen (1978). Many formats have been proposed to assess speaking ability. Some of these include pictures to elicit speech, reading short sentences, and assigning a topic to elicit a sustained speech.

The last point that needs to be considered in developing a language test is how it should be administered. Many instructions for English-language proficiency tests are given in English. The logic behind this is that if a person knows enough English to take the test, that person should be able to understand the instructions in English. Both the Basic and Intermediate ELTs' directions are given in the native language. This will be continued for the revised ELTs. However, Ramos (1981) showed that when instructions for a test were given in the native language of the person taking the test, significant gains in scores resulted. The effects of this on test validity for educational or job success criteria are not known.

The Basic and Intermediate tests were revised by first generating 120 items for each test. Second, the item pools were administered along with the existing tests to native English-speakers to ensure all ELT items tested only English proficiency and not specialized knowledge or other extraneous factors. Next, pretesting with the ELTs occurred on a small group of foreign students to ensure that items discriminated among ability levels of non-English-speakers. Finally, a field test was conducted on foreign employees for final item selection.

## II. TEST CONSTRUCTION

### Basic English Language Test

The intent of the revision of the Basic ELT was to increase the number of items in each test from 20 to 25 but to allow the content to remain unchanged. This would allow easier test score interpretations (total score of 100 instead of 80), and it would increase test reliability. Therefore, 120 new items were generated for each test that were similar in nature to those in the existing tests.

The first step was to categorize the existing tests into some meaningful context. The correct response to each item was assigned a word frequency according to Carroll, Davies, and Richman (1971). These frequencies were categorized according to the three broad frequency categories established by Lorge and Thorndike (1944). These categories were at least 100 occurrences per million, at least 50 occurrences per million, and less than 50 occurrences per million. New items were chosen for each test according to the same proportion of difficulty as appeared in the original versions of the Basic ELT. Lists of 120 new items per test were then presented to the Aerospace Medical Division's Medical Illustration Section for graphic artwork.

Next, distractors were generated for the two multiple-choice tests (Listening and Reading). Listening test distractors were derived by cross-cultural phonetic similarities (e.g., "chicken" [Spanish=pollo] distracting the word "pole"), by vowel contrasts (e.g., "ship" distracting "sheep"), and by grammar (e.g., "house dog" distracting "dog house"). Reading test distractors were created with spelling distractors (e.g., "bazball" distracting "baseball") and similar-appearing English words (e.g., "army" distracting "arm"). No distractors were necessary for the Writing and Speaking tests by their nature.

### Intermediate English Language Test

In contrast to the Basic ELT, a complete revision was necessary for the Intermediate ELT. A 100-point battery that was content-parallel to the Basic ELT was required. Although the existing Intermediate ELT contained three sections, it essentially measured only reading ability. The new Intermediate ELT was constructed to measure writing, listening, reading, and speaking abilities.

According to Lado (1961), writing a language consists of knowing the language's rules for grammar, vocabulary, spelling, and punctuation. Assessing writing skill is less a matter of sampling the act of physically writing words and sentences and more a matter of testing one's knowledge of a language's writing rules. Therefore, for the Writing Test, 120 multiple-choice items were developed that were equally divided among testing rules for grammar, vocabulary, spelling, and punctuation. Distractors were chosen according to the rule being tested (e.g., grammar--went, gone; vocabulary--lake, sea, ocean; spelling--light, lite; and punctuation--!, ?).

Listening test items were constructed with an aural English lead sentence and four English sentences from which the test-taker must choose the most similar to the lead sentence in meaning. The leads were all free utterance which can appear independently in conversations. Care was taken to avoid technical material and to limit the leads to only one sentence. These restrictions ensured that the content of the lead material was equally familiar (or unfamiliar) to all test-takers. Distractors were selected primarily to determine whether the test-takers understood the meaning of the leads. The distractors explored grammatical and/or syntactical structure (e.g., "bicycle between two cars" versus "car between two bicycles") and vocabulary (e.g., "equal" versus "different"). One hundred twenty multiple-choice items were developed.

The third multiple-choice test of the revised Intermediate ELT is the Reading test. This test uses the CLOZE procedure described in the Introduction. The passages were taken from discarded items of an Armed Services Vocational Aptitude Battery (ASVAB) updating effort. The ASVAB is an aptitude test battery used by all of the Armed Services to select and classify enlisted personnel. According to the FORCAST method of determining reading grade level (RGL), which was developed by Caylor, Sticht, Fox, & Ford (1973), these passages had a mean RGL of 10.78. Every seventh word was deleted from these passages. The only exceptions were the first and last sentences, which were left intact to provide an understandable context for the passage. The 120 deleted words became the correct answers. Development of distractors varied according to the target answer. Verbs and adverbs generally tested past tense and plurals (e.g., "is," "was," "are," "were"). Noun distractors made sense in the sentence but not in the context of the passage. Adjectives tended toward opposites (e.g., "hot," "cold") whereas combinations of distractors were used for conjunctions (e.g., "and," "or," "where"). Thus, no single set of rules was used to develop distractors, but they were selected according to how plausible they were in the context of the passage.

The Speaking test was adapted from the paradigm advocated by Mullen (1978). In this test, two raters carry on a 15-minute conversation with the test-taker. After 15 minutes (in practice, 10 minutes was found to be sufficient), the two judges rate the individual's vocabulary, pronunciation, fluency, grammar, and overall oral proficiency, based upon behaviorally anchored rating scales. An example of the rating sheet is provided in Appendix A. Each scale ranges from Poor to Excellent; with Poor = 1, Marginal = 2, Fair = 3, Good = 4, and Excellent = 5. Thus, with five scales and a maximum of five points per scale, a total maximum score of 25 is possible on the Speaking test. Twenty-five points was targeted to be the maximum score on each test. This would yield a 100-point battery, which would parallel the Basic ELT.

## III. ITEM SELECTION METHOD

The overall plan for item selection and test validation called for three phases which included administering the ELTs to native English-speakers, screening on a small group of foreign students, and field testing with foreign nationals already working at bases overseas. Trying out the revised Basic and Intermediate ELTs on English-speakers was necessary to detect any extraneous factors in them, such as testing memory, intelligence, or technical matter. The rationale for screening the ELTs on a small group of foreign students prior to field testing was twofold. First, screening the ELTs provided evidence of whether the ELTs could discriminate among foreigners as well as do other current testing instruments. Secondly, screening the ELTs allowed a reduced item pool to be field tested. Final item selection was based upon the results of the field test.

As mentioned above, the first phase entailed administering the Basic and Intermediate ELTs to a native English-speaking group. It was first necessary to identify a sample of "average" English-speakers. A random sample of Air Force basic trainees was selected for this purpose. For the Basic ELT, a sample of 348 trainees were used, of which 66% were high school graduates, 76% were males, and 66% were less than 21 years old. The Intermediate ELT sample was composed of 635 basic trainees, of which 80% were high school graduates, 76% were males, and 74% were less than 21 years of age. All 120 items on each subtest of the Basic ELT item pool were administered to the former sample. The 120 items in each subtest of the Intermediate ELT item pool were administered to the latter sample, along with the existing Intermediate ELT, in a counterbalanced design. Any extraneous factors in the final items were avoided by eliminating items missed by more than 75% of the basic trainees or items that showed significantly positive distractor biserials.

The next phase of this project pretested both the existing and revised item pools on a group of foreign students. Arrangements were made with the Defense Language Institute (DLI) to utilize a sample of their students, who already had scores on the English Comprehension Level (ECL) examination. The ECL is a test used by the Department of Defense to measure the English proficiency of foreigners who receive U. S. military training. These scores would be used as a measure of concurrent validity. The Basic ELT sample consisted of 99 students, of whom 90% had 12 years or more of education, 100% were male, and 72% were less than 28 years old. The Intermediate ELT sample contained 99 students, of whom 90% had 12 or more years of education, 99% were male, and 54% were less than 28 years of age. The existing ELTs and replacement item pools were administered to the samples in a counterbalanced design. The results were used to make the final item selection for the Basic ELT and to reduce the Intermediate ELT item pool to 60 items per subtest for field testing.

The last phase of this project involved field testing the ELTs with foreign nationals currently working at bases overseas. Because the format of the Basic ELT was essentially unchanged, the new tests were field tested only on 17 employees at Howard Air Force Base (AFB), Panama. This was done to ensure the Basic ELT could discriminate among foreign national employees. The major thrust of the field testing centered on the Intermediate ELTs. The item pools were administered to 490 foreign national employees randomly selected at 16 bases overseas. The following nationalities were included in the field test: German, Portuguese, Italian, Spanish, Turkish, Greek, Filipino, and Korean. Eighty-four percent of the sample had at least 9 years of education and were at least 25 years old; 44% were male. In addition to administering the Intermediate ELT item pools, a supervisor's rating sheet was distributed to each subject's work supervisor. This supervisor's rating sheet gave a measure of the Intermediate ELT's validity. Appendix B shows an example of the rating sheet.

## IV. RESULTS

When the results from pretesting the Basic ELT on basic trainees were analyzed, the mean score (on 120-item tests) were: for the Reading test, 115.41; for the Writing test, 114.91; and for the Listening test, 116.72. Scoring for the Speaking test is on a nominal scale and, as expected, the ratings' mode was "no detectable accent." Pretesting the Intermediate ELTs on basic trainees provided similar results. Mean scores on each test were: Reading test, 102.94; Writing test, 109.63; and Listening test, 115.13. Since only five items (of 360) on the Basic ELT and only 30 items (of 360) on the Intermediate ELT failed to reach the .75 difficulty level and none had significant positive distractor biserials, all of the items were presented to the DLI students in the next phase. These 35 unacceptable items were subsequently eliminated.

When the replacement item pools were administered to the foreign students at DLI, lower scores were observed on all tests than were found with the basic trainees. Mean scores (of 120 items) on the Basic ELT were: Reading test, 78.99; Writing test, 79.75; and Listening test, 94.71. Each Basic Speaking test item's ratings were normally distributed. Final items were selected by comparing the existing Speaking test item distributions with those of the replacement item pool distributions. The criteria used for selection were similarity to the existing Speaking test item difficulty level and the ability of the item to discriminate (i.e., having a relatively normal distribution). Mean Intermediate ELT scores obtained by foreign nationals were also lower than those of the basic trainees: Reading test, 78.81; Writing test, 81.81; and Listening test, 87.27. As shown in Tables 2 and 3, the item pools selected for field testing showed significant positive correlations with both forms of the existing ELTs and DLI's ECL examination.

Table 2. Basic ELT Correlations on DLI Students
(N = 99)

| | | Field test item pools (60 items/test) | | |
|---|---|---|---|---|
| Existing tests | Items | Reading | Writing | Listening |
| Reading-Form A | 20 | .68 | | |
| Reading-Form B | 20 | .64 | | |
| Writing-Form A | 20 | | .84 | |
| Writing-Form B | 20 | | .84 | |
| Listening-Form A | 20 | | | .55 |
| Listening-Form B | 20 | | | .58 |
| ECL examination | | .72 | .79 | .74 |

Note. All correlations were significant at the .01 level.

Table 3. Intermediate ELT Correlations on DLI Students
(N = 99)

| | | Field test item pools (60 items/test) | | |
|---|---|---|---|---|
| Existing tests | Items | Reading | Writing | Listening |
| Part I Form A | 30 | .70 | .65 | .67 |
| Part I Form B | 30 | .64 | .58 | .56 |
| Part II Form A | 27 | .55 | .45 | .53 |
| Part II Form B | 27 | .60 | .55 | .63 |
| Part III Form A | 23 | .55 | .41 | .59 |
| Part III Form B | 23 | .56 | .52 | .57 |
| ECL examination | | .68 | .59 | .82 |

Note. All correlations were significant at the .01 level.

A comparison of difficulty levels was made between the Basic and Intermediate ELTs using the data obtained from the DLI students. Since all students were tested on the ECL examination, mean ELT scores were generated at various ECL score intervals. For example, students who scored between 41 and 50 on the ECL had mean Basic ELT scores as follows: Listening - 40.75, Reading - 36.50, and Writing - 28.00. In the same ECL score range, students' Intermediate ELT scores were the following: Listening - 20.60, Reading - 31.40, and Writing - 25.00. Although these data should be viewed with caution due to the small sample cell sizes, it can be concluded that the Intermediate ELT is more difficult than the Basic ELT.

The third and final phase of this project was the field test. The Basic Reading test scores ranged from 13 to 49, with a mean of 33.59; the Writing test score mean was 23.18, with a range of 7 to 50; and the Listening test scores ranged from 17 to 48, with a mean of 34.29. These tests had a maximum score of 50. Results of the Speaking test revealed similar findings to those for the DLI sample: good discrimination between high and low ability. When the test reliabilities (Reading = .98, Writing = .98, and Listening = .92) obtained from the DLI sample were considered along with the range of scores obtained in the field test, the Basic ELT showed that it could discriminate among individuals in the Howard AF8 sample.

As mentioned previously, the Intermediate ELT underwent a major revision. Therefore, the field testing was much more extensive for the Intermediate than the Basic ELT. The mean score for the Writing test score (of a possible 60) was 49.26, standard deviation was 9.37, and test

reliability was .93. Mean Listening test score was 46.15, standard deviation was 12.92, and test reliability was .96. The mean of the Reading test was 44.71, the standard deviation was 11.50, and test reliability was .94. Using the scoring method described in the Test Construction section, the mean Speaking test score was 20.24 (of a possible 25 points), with a standard deviation of 3.99, and an interrater reliability correlation of .87. Intercorrelations of the four tests and the supervisors' ratings are shown in Table 4. These correlations reveal positive signifirant relationships among the Intermediate ELT tests and the supervisors' ratings.

Table 4. Intercorrelations of the Parts of the Intermediate ELTs and Supervisors' Ratings (N = 490)

| | Writing test | Listening test | Reading test | Speaking test |
|---|---|---|---|---|
| Listening test | .80 | | | |
| Reading test | .84 | .86 | | |
| Speaking test | .46 | .62 | .54 | |
| Supervisors' Ratings | .41 | .49 | .47 | .57 |

Note. All correlations were significant at the .01 level.

The final score for the Intermediate ELT is obtained by summing the four test scores. Using the supervisors' ratings as a measure of validity, a correlation of .52 was found for this summed score. This is lower than the .57 for the Speaking test and is somewhat surprising. The cause for the drop in the validity coefficients is likely due to the lower variance of the Speaking test in relation to the variances of the other three tests. If the individual tests could be equally weighted in operation, higher validity would result. For example, by unit weighting the Writing, Listening, and Reading tests and applying a weight of 3 to the Speaking test, the validity is increased to .56.

Other than creating associated materials for the ELTs, such as administration manuals and scoring keys, the final task of this project was to separate the field test item pools into two operational versions. Information obtained from the DLI students in Phase 2 was used as a basis for separating the items in the Basic ELT. Each item's level of difficulty was matched with another's difficulty level to be placed in one of two alternate forms. This method resulted in the following mean levels of item difficulty for each form on each test: Writing test = .59, Listening test = .69, Reading test = .69, and Speaking test = 2.20.

The rationale for assigning items to Forms A and B of the Intermediate ELT was based on data from the Phase 3 field test. Only 50 out of 60 items per test were needed from the field test item pool. The statistically least powerful items were discarded. That is, items with positive distractor biserials or items above the .92 level of difficulty were not selected to be included in the final test forms. The remaining 50 items were then divided into two forms of 25 items each, based on their item difficulties. The following were the mean levels of item difficulty for both forms of each test: Writing test = .80, Listening test = .75, and Reading test = .75. Based on the field test sample, the correlations between the individual test forms were .85 for the Writing test, .91 for the Listening test, and .85 for the Reading test. According to the Wherry and Gaylord (1943) estimate of reliability, the reliability for the composite of all subtests of the Intermediate ELT was .96. Appendix C gives a summary of the statistics on the final versions of the Basic and Intermediate ELTs.

## V. RECOMMENDATIONS

From the data generated by this effort, it is concluded that two equivalent forms of the Basic and Intermediate ELTs have been generated. Furthermore, based upon comparisons with the ECL and existing ELTs, the new ELTs measure a person's command of English as a second language. Therefore, it is recommended that the new Basic and Intermediate ELTs be implemented.

Interpretations of test scores could be enhanced by future research. It was not feasible to collect data on a sample sufficiently large nor representative of all worldwide applicants who normally take the Basic and Intermediate ELTs. These tests could be adequately normed by collecting test scores and demographic information on individuals who apply for work at bases overseas and take the new ELTs. By doing this, separate norms could be established for each language group. Also, these data could be used as a basis to decide whether to administer the Basic ELTs or to administer the Intermediate ELTs. This would be accomplished by establishing appropriate difficulty ranges for various ability levels.

## REFERENCES


Buros, O. (1978). The eighth mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press.

Carroll, J., Davies, P., & Richman, B. (1971). The American heritage word frequency book. Boston; Houghton Mifflin.

Caylor, J., Sticht, T., Fox, L., & Ford, J. (1973). Methodologies for determining reading requirements of military occupational specialties (HumRRO-TR-73-5). Alexandria, VA: Human Resources Research Organization.

Corbett, P. (1980, June). Setting standards in English language. Paper presented at the International Symposium on Educational Testing. (ERIC Document Reproduction Service No. ED 198 726)

Hisama, K. (1977a). Design and empirical validation of the cloze procedure for measuring language proficiency of non-native speakers. Dissertation Abstracts International, 37(9-A), 5766A.

Hisama, K. (1977b, April). Patterns of various ESOL proficiency test scores by native language and proficiency levels. Occasional papers on linguistics, No. 1. Proceedings of the International Conference on Frontiers in Language Proficiency and Dominance Testing. (ERIC Document Reproduction Service No. ED 144 409)

Lado, R. (1961). Language testing. London: Longmans.

Lombardo, M. (1981). The construction and validation of listening and reading components of the English as a Second Language Assessment Battery. Published as part of the Ethnoperspectives Project. (ERIC Document Reproduction Service No. ED 212 155)

Lorge, I., & Thorndike, E. (1944). The teacher's word book of 30,000 words. NY: Teachers' College, Columbia University.

Mullen, K. (1978). Direct evaluation of second language proficiency: The effect of rater and scale in oral interviews. Language Learning, 28(2), 301-308.

Pike, L. (1979). An evaluation of alternative item formats for testing English as a foreign language. Princeton, NJ: Educational Testing Service.

Ramos, R. (1981). Employment battery performance of Hispanic applicants as a function of English or Spanish test instructions. Journal of Applied Psychology, 66(3), 291-295.

Scholz, G., & Scholz, C. (1981). Multiple choice cloze tests of ESL discourse: An exploration. Paper presented at the annual TESOL convention. (ERIC Document Reproduction Service No. 208 656)

Schulz, R. (1977). Discrete-point vs. simulated communication testing in foreign languages. Modern Language Journal, 61(3), 94-101.

Stubbs, J., & Tucker, R. (1974). The cloze test as a measure of English proficiency. Modern Language Journal, 58(5-6), 239-241.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 415-433.

Wherry, R., & Gaylord, R. (1943). The concept of test and item reliability in relation to factor pattern. Psychometrika, 8, 247-264.

APPENDIX A: SPEAKING TEST RATING SHEET

# SPEAKING TEST RATING SHEET

Name_____ Employee ID Number_____

After the person being rated has been dismissed, circle either excellent, good, fair, marginal, or poor on each of the five rating scales.

## Vocabulary

Excellent - Uses a large number/variety of words correctly.
Good - Only occasionally uses a word incorrectly or has difficulty choosing a word.
Fair - Often has difficulty choosing an appropriate word.
Marginal - Great difficulty using words other than the most simple.
Poor - Is not able to express even a simple sentence.

## Pronunciation

Excellent - Few, if any, traces of accent.
Good - Alway understandable, but definite accent.
Fair - Heavy accent causes occasional misunderstandings.
Marginal - Very heavy accent, repetition necessary to convey meaning.
Poor - Accent causes speech to be barely understood.

## Fluency

Excellent - Smooth and effortless speech.
Good - Speaks readily with only occasional hesitation.
Fair - Falters and hesitates often, pauses are frequent but usually short.
Marginal - Usually hesitant speech, sometimes forced into silence.
Poor - Halting and fragmentary speech, conversation virtually impossible.

## Grammar

Excellent - Few, if any, grammar or word order problems.
Good - Occasional grammar or word order problems.
Fair - Errors often cause meaning of sentences to become obscured.
Marginal - Great difficulty using correct grammar or word order, frequently uses incorrect verb tense, nouns, adjectives, etc.
Poor - Speaking can't be understood due to grammar errors.

Overall Oral Proficiency - Basing your decision on all of the above criteria, rate the examinee on his or her overall command of the English language.

Excellent
Good
Fair
Marginal
Poor

APPENDIX B: SUPERVISOR'S RATING SHEET

## SUPERVISOR'S RATING SHEET

First, print the employee's name and identifying number in the spaces provided. Then, as objectively as you can, rate the employee using the following eight scales. Simply circle either excellent, good, fair, marginal, poor, or not observed on each of the scales. Please rate the individual on all rating scales.

Name_____ Employee ID Number_____

### 1. Vocabulary

Excellent - Uses a large number/variety of words correctly.
Good - Only occasionally uses a word incorrectly or has difficulty choosing a word.
Fair - Often has difficulty choosing an appropriate word.
Marginal - Great difficulty using words other than the most simple.
Poor - Is not able to express even a simple sentence.
Not observed

### 2. Punctuation and Spelling

Excellent - Writing has virtually no punctuation or spelling errors.
Good - Makes occasional punctuation or spelling errors.
Fair - Frequent errors cause writing to be difficult to read.
Marginal - Many errors cause writing to be very difficult to read.
Poor - Extreme number of errors cause writing to be misunderstood.
Not observed

### 3. Grammar

Excellent - Few, if any, grammar or word order problems.
Good - Occasional grammar or word order problems.
Fair - Errors often cause meaning of sentences to become obscured.
Marginal - Great difficulty using correct grammar or word order, frequently uses incorrect verb tense, nouns, adjectives, etc.
Poor - Writing and speaking can't be understood due to grammar errors.
Not observed

### 4. Fluency

Excellent - Smooth and effortless speech.
Good - Speaks readily with only occasional hesitation.
Fair - Falters and hesitates often, pauses are frequent but usually short.
Marginal - Usually hesitant speech, sometimes forced into silence.
Poor - Halting and fragmentary speech, conversation virtually impossible.
Not observed

5. Pronunciation

Excellent - Few, if any, traces of accent.
Good - Always understandable, but definite accent.
Fair - Heavy accent causes occasional misunderstanding.
Marginal - Very heavy accent, repetition necessary to convey meaning.
Poor - Accent causes speech to be barely understood.
Not observed

6. Reading Comprehension

Excellent - Can read virtually any English word.
Good - Has some difficulty recognizing some English words.
Fair - Does not recognize many English words.
Marginal - Can read only simple English words.
Poor - Cannot understand most English words.
Not observed

7. Listening Comprehension

Excellent - Can understand oral instructions with no misunderstandings.
Good - Sometimes needs oral instructions repeated to understand what is being said.
Fair - Often misinterprets oral instructions, several repetitions sometimes necessary.
Marginal - Can only understand simple oral instructions, errors often occur.
Poor - Seldom understands oral instructions.
Not observed

8. Ability to perform job based on English proficiency

Excellent - Use of English does not impair job performance.
Good - English usage slightly affects employee's job performance.
Fair - Job performance is frequently hindered by use of English.
Marginal - Use of English severely affects job performance.
Poor - Lack of English skills causes job performance to be accomplished incorrectly most of the time.
Not observed

APPENDIX C: REVISED ELTs' DESCRIPTIONS

Table C-1. Basic ELTs Statistics

| Test | Items | Mean difficulty | Reliability | A-B correlation |
|---|---|---|---|---|
| Writing | 25 | 14.75 | .98 | .95 |
| Listening | 25 | 17.25 | .92 | .89 |
| Reading | 25 | 17.25 | .98 | .94 |
| Speaking | 25 | 13.75 | .91 | .85 |

Note. These data are based upon the DLI sample (N = 99).

Table C-2. Intermediate ELTs Statistics

| Test | Items | Mean difficulty | Reliability | A-B correlation |
|---|---|---|---|---|
| Writing | 25 | 20.07 | .93 | .85 |
| Listening | 25 | 18.80 | .96 | .91 |
| Reading | 25 | 18.68 | .94 | .85 |
| Speaking | N/A | 20.28 | .87 | N/A |

Note. These data are based upon the overseas field test sample (N = 489).

U.S. GOVERNMENT PRINTING OFFICE:1986-761-057/40036