DOCUMENT RESUME

ED 206 627 TM 009 602

AUTHOR Pike, Lewis W.
TITLE An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO TOEFL-RR-2
PUB DATE Jun 79
NOTE 110p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Cloze Procedure; *English (Second Language); Essay Tests; Foreign Students; Higher Education; Interviews; *Language Proficiency; *Language Tests; Multiple Choice Tests; Scoring; *Test Format; Test Reviews; *Test Validity
IDENTIFIERS *Test of English as a Foreign Language; *Test Revision

ABSTRACT

This evaluative and developmental study was undertaken between 1972-74 to determine the effectiveness of items used for the Test of English as a Foreign Language (TOEFL) in relationship to other item types used in assessing English proficiency, and to recommend possible changes in TOEFL content and format. TOEFL was developed to assess the English proficiency of non-native English-speaking students applying to institutions of higher education in the United States. Questions of validation, criterion selection and content specification were first investigated before nine written and oral TOEFL item formats were evaluated for possible use in a revised test. Both original and new formats were administered to 98 Peruvian, 145 Chilean and 199 Japanese subjects in their native countries. Open ended response measures and multiple choice measures were examined. Intercorrelations among test scores indicated that the test could be revised to incorporate three instead of five components: (1) listening comprehension; (2) English structure and writing ability; (3) reading comprehension and vocabulary in context. Four objective subtests aimed at increasing TOEFL effectiveness, and tailored criterion measures of English productive skills, speaking and writing were also developed. (AEF)

# TOEFL Research Reports

REPORT 2
JUNE 1979



## AN EVALUATION OF ALTERNATIVE ITEM FORMATS FOR TESTING ENGLISH AS A FOREIGN LANGUAGE

Lewis W. Pike
Educational Testing Service

ACKNOWLEDGMENTS

An investigation of this magnitude requires the efforts of many individuals other than the principal investigator. Although the author cannot hope to acknowledge his debt to everyone who helped, he wishes to express special thanks to individuals and institutions whose contributions were particularly great.

Thanks are owed to the past and present members of the TOEFL Committee of Examiners and to the ETS staff working with the TOEFL program, whose suggestions before and during the investigation helped to shape the study and the ideas embodied in it.

Data collection was especially arduous, requiring over eight hours of testing time for each subject, and innumerable hours of work by collaborators responsible for their recruitment and testing. The testing at each site was on a large scale and included, in addition to group testing, the conducting of an individual tape-recorded interview with each participating student. This required extensive coordinating by the overseas collaborators as well as extended use of the facilities and support staffs of their respective institutions. The collaborators and their institutions included the following:

| | |
|---|---|
| Lima, Peru | Professor Pablo Willstätter G., Vice President of Education, and Frank Valcárcel, Educational Assistant, Instituto Peruano de Fomento Educativo. |
| Santiago, Chile | Professor Jack R. Ewer, Chairman, Department of English, Faculty of Philosophy and Education, and Professor Evan Hughes-Davies, Coordinator of Language Programs, Faculty of Physical Sciences and Mathematics, University of Chile. |
| Tokyo, Japan | Dr. Floyd M. Cammack, President, Language Education Associates, Incorporated, and Dr. Mitsuo Hashimoto, Academic Director, The International Education Center. |

Special thanks are due Fred Godshalk, who served as Chief Reader, to Roberta Wakefield and her Essay Reading Group, and to the other essay readers, for their work in judging the quality of the essays. Thanks are also due Richard Yurko, Nancy Kuykendall, and the other interview judges for their work in judging participating students' English speaking skills, on the basis of tape-recorded interviews.

Appreciation is extended to Richard Reilly of ETS for his assistance in obtaining the Japanese data, and for his advice on certain aspects of the data analysis, and to Miriam Godshalk for her invaluable assistance in writing this report.

Finally, the author wishes to express his appreciation for the financial support provided by the TOEFL program to carry out the study.

The Test of English as a Foreign Language (TOEFL) is well known among university officials and others who are concerned with the admission of foreign students who are nonnative speakers of English to institutions of higher education in the United States and Canada. It is certainly well known among the thousands of foreign students who take it each year, in many foreign countries, as one of the requirements for entrance to American colleges and universities. Some of them may think of it as a "devil" of a test--for the acronym is all too close to the German word for devil!

All these people, the foreign students and the university officials, have a right to expect that the TOEFL is the fairest, most accurate, and most valid test of its type that can be devised.

Constructing such a test is not easy. Questions have frequently been raised as to whether the TOEFL is in act as good as it might be. Because the test has remained in essentially the same form over many years, some people may have arrived at the impression that Educational Testing Service--the organization responsible for the test--is resistant to making changes in it.

The present monograph will give the lie to such an impression. It reports an extensive research study that was designed to explore possible changes in the format and content of the TOEFL. It illustrates several points: that it is extremely difficult and expensive to conduct a really thoroughgoing study of possible changes; that many apparently reasonable suggestions for change turn out not to be so valid and feasible after all; and that nevertheless, some changes

prove to be promising, desirable, and feasible. Most of all, however, it indicates, at least to me, that the essential philosophy and direction of the TOEFL as it now exists, or as it might be modified in certain ways suggested here, is sound and credible.

Teachers of English (and perhaps teachers of other languages), as well as language testing specialists, will find much of interest and value in this monograph. I welcome Dr. Pike's monograph as a substantial contribution to the field.

John B. Carroll
University of North Carolina
at Chapel Hill

CONTENTS

LIST OF TABLES

## PURPOSE AND BACKGROUND

A major purpose of the Test of English as a Foreign Language (TOEFL) is to provide information useful to colleges and universities in making decisions regarding the admission, placement, and possible assignment to special language instruction of foreign students planning to study in the United States and Canada. Its unique role is to assess the foreign student's competence in English that he will need in order to successfully pursue a program of studies at a college or university where English is the medium of instruction.

The overall purpose of the present study was to obtain information useful for evaluating and revising TOEFL content and content specifications. To achieve this purpose, questions of validation, criterion selection, and content specifications were investigated.

The tasks involved in carrying out this study were developmental as well as evaluative. Thus, four objective subtests that might increase the effectiveness of TOEFL, and tailored criterion measures of English productive skills, speaking and writing, were developed. Also included in the study were two open-ended response measures which have shown particular promise for testing English as a foreign language. One is a rewriting task used by Kellogg Hunt (1970a, 1970b), and the other is the "Cloze procedure" task (Taylor, 1953), in which subjects are instructed to replace words that have been deleted from prose passages.

Validation, criterion selection, content specifications, and related questions regarding TOEFL will be considered first. The subtests used in carrying out the study and the data resulting from the study will be presented in later sections of this paper.

## Questions of Validity and Criterion Selection

In an evaluation of a test or its components for possible revision, consideration must be given to questions of validity, reliability, and practicality. In keeping with the basic purpose of this study, all three were considered, with special emphasis on the question of validity. Tied to the question of validation is that of criterion selection or development. In the present study, criterion selection is, in turn, based on the purpose and content of the TOEFL examination, and on the goal of the validation study itself.

### English Language Skills as Criteria

Although TOEFL is used as an admissions test, it, unlike other admissions tests such as the Scholastic Aptitude Test and the Graduate Record Examination, is not intended to serve as a predictor of academic success, particularly as this is measured by grade-point average. Rather, TOEFL is used to ascertain if foreign students have sufficient command of English to enable them to study at institutions where English is the medium of instruction without being handicapped by inadequate communication skills. The first implication of the above considerations is that it is more appropriate to use measures of various areas of competence in English as criteria rather than some more inclusive index such as grade-point average in establishing the validity of TOEFL. A second implication of viewing the essential task of TOEFL as the assessment of current English language skills, rather than as the prediction of future academic success, is that the validation procedure becomes one of concurrent instead of predictive validity.

At the time this study was conducted, the TOEFL examination was made up of the following five sections, each directed to one of the areas of competence in English as a second language: I. Listening Comprehension; II. English Structure; III. Vocabulary; IV. Reading Comprehension; and V. Writing Ability. The need for such differentiated test of second-language performance was described by Carroll (1968):

> The problem of different areas of competence becomes much more acute in dealing with a second or foreign language, where the experiences of learners are likely not to be as homogeneous as those of native language learners, and where various well-known difficulties interpose themselves in learning--the interference of the native language, the slow progress due to the student's lack of time or motivation for study, etc. Furthermore, since learning a second language often makes much more use of written material than does learning a native language (where reading is rarely started until the spoken language is fairly well mastered), competence in spoken and written aspects may develop somewhat independently, and these competences must be separately assessed. It is also much more important to observe the distinction between productive and receptive skills because progress in these two aspects may not proceed pari passu as it ordinarily does in the native language. It is quite possible for a competence to relate specifically to production and not to reception, or vice versa [p. 52].

Another partitioning of skills, cutting across the competencies of listening and reading and of speaking and writing, is that between vocabulary and structure. Of these six areas of English language competence, the TOEFL examination provides direct assessment of the receptive abilities, listening and reading, and of vocabulary and English structure, as well as indirect assessment of writing competence. Because of feasibility constraints, English speaking skills are not measured. Thus the TOEFL examination provided five scores considered potentially relevant for diagnostic as well as admissions interpretation.

The differentiated nature of the TOEFL examination and the purpose of the present validation study further define the question of criterion selection. This study is not designed to compare TOEFL with other

examinations, nor is it limited to obtaining an index of how well it measures overall performance in English as a second language. Rather, it is intended to indicate how well several component language competencies (including the speaking of English) are estimated by the TOEFL subtests and by the alternative experimental measures developed for this study. Thus, instead of one overall measure of English competence, a set of criterion measures directed to specific component skills is called for.

The above discussion provides a rationale whereby the present study must involve concurrent validation, directed to six areas of English language competence. Before discussing criterion selection for these six areas, it will be helpful to review previous validation studies involving TOEFL, and to make certain theoretical observations about the problem of validation.

Previous TOEFL Validation Studies

A summary of predictive and concurrent validity studies involving TOEFL is provided in the booklet Test of English as a Foreign Language: Interpretive Information (1970). The predictive validity studies, with their emphasis on the grade-point average criterion, provide little information bearing on the evaluation and revision of TOEFL content specifications. More to the point are the concurrent validity studies summarized in the booklet, although most of these are focused on comparing TOEFL scores with scores on similar kinds of tests, such as the American Language Institute Test of Proficiency in English and the Michigan Test of English Language Proficiency. Among the concurrent validity studies summarized in the booklet, a study, carried out by the staff of the American Language Institute at Georgetown University in

cooperation with the staff of Educational Testing Service (Pitcher and Ra, 1967), is the most relevant to the questions posed in this paper. In that study, correlations were obtained between each TOEFL subtest and a criterion of judged essay-writing performance. The correlations with the criterion were generally in the order one would logically expect, with Writing Ability and English Structure correlating highest (.74 and .74) and Listening Comprehension lowest (.56).

Criterion Selection

In most validation studies, the problem of criterion selection and development cannot be fully resolved. Even when ultimate criteria can be agreed upon and clearly stated, feasibility constraints typically require compromises of such nature that the criterion measures adopted are only relatively more direct than those being validated. Nevertheless, the reduced constraints of cost and time in an experimental study allow the use of criteria that can measure the target abilities more directly than those being validated, and that may have, as well, greater face validity. Furthermore, the circularity of using tests to validate tests may be partly offset by a consideration of construct validity. For example, the Pitcher and Ra finding that Writing Ability correlated higher with the essay-writing criterion than did most of the other TOEFL sections lends credence to its construct validity as a writing measure beyond that implied by the size of the correlation itself.

The receptive language skills of listening and reading appear to be quite directly measured by the Listening Comprehension and Reading Comprehension sections of TOEFL. Although more direct measures of these skills which could serve as criteria can be readily conceived (using videotaped

classroom sessions to test listening comprehension, for example), they were not considered feasible within the constraints and scope of the present study. Therefore, no criterion measures were developed for listening and reading. For the productive language skills of speaking and writing, however, criterion measures were developed that called for actual performance in speaking and writing. In the spoken mode, English structure and vocabulary criteria were provided through judgments of the tape-recorded interviews used to assess speaking ability, along specified dimensions.

The Hunt rewriting task and the Cloze procedure tasks illustrate the relative nature of the question, "What is a criterion?" In the present study, each may be considered a quasi-criterion that was used to validate less direct, multiple-choice measures but was itself validated against even more direct measures of English language competence.

## Questions of Content Specifications

The development of alternative TOEFL measures and the evaluation of these and of present TOEFL subtests were guided by questions concerning the content specifications. Some of these questions have been raised by the TOEFL Committee of Examiners and ETS staff members at various times since the inception of TOEFL; others derive from a close examination of TOEFL and from reviewing the relevant literature of language testing.

For convenience, questions regarding the present TOEFL subtests will be presented first, followed by those related to the open-ended Hunt and Cloze procedure tasks and the alternative multiple-choice measures. Directions and sample items for each of these measures are provided in the Plan and Procedure section of this report.

Present TOEFL Sections

In general, there is an interest in the reliability, efficiency, and concurrent validity of each of the TOEFL sections, and in the degree of construct validity indicated by relationships within the full set of estimator and criterion variables. For the TOEFL sections having two or three parts, each with a different item format, the question of the relative merits of each part is also of interest. In addition to these general questions, questions and observations that emerged regarding the individual TOEFL sections are given below.

T1, Listening Comprehension. Although the background information and questions in this section are presented aurally, the answer choices are given in written form only. The influence of the reading component on the resulting listening scores is, therefore, of interest.

T2, English Structure. The items in this section seem to stress standard English usage as it is spoken in the United States more than differences in meaning that are conveyed by structure. Generally, however, this section appears well accepted because the importance of testing English structure is often emphasized in discussions of testing English as a foreign language (Bilyeu, 1969; Carroll, 1968; Fisher and Masia, 1965).

T3, Vocabulary. There has been a general concern that vocabulary may be given too much emphasis as compared with English structure. At various times, one or more TOEFL Committee or ETS staff members have suggested dropping the Vocabulary section altogether. One criticism is that a vocabulary section may encourage an overemphasis on vocabulary scores; this probably contributes little that is unique to what is measured by the other TOEFL sections. Yet another criticism is that,

too often, the vocabulary items include low-frequency, esoteric words that are of little practical use to the foreign student.

T4, Reading Comprehension. This section has high face validity, but it requires substantially more testing time (and test development costs) to reach a given level of reliability than do most other sections of TOEFL. The general question, of course, is whether a more efficient measure can be developed to replace all or part of this section, with its format of several reading passages, each followed by a number of questions.

T5, Writing Ability. The value of this section has been questioned by the TOEFL Committee and ETS staff, some of whom have suggested dropping it unless validity data clearly indicate that it should be retained. The use of a writing sample has been suggested as a replacement for the less direct Writing Ability section. Thus, the question of how well Writing Ability scores estimate essay-writing criterion scores is of particular interest.

Hunt and Cloze Procedure Tasks

Hunt rewriting task. The principal measure obtained from Kellogg Hunt's rewriting task, "Words per T-Unit," is essentially a refinement of the familiar sentence-length measure. The Words per T-Unit measure has worked very well for estimating the English language "syntactic maturity" of American children and adults (Hunt, 1970b), and it has intriguing possibilities for doing the same with respect to foreign students' command of the sentence-embedding aspect of English structure (Hunt, 1970a).

One question of interest is whether the Words per T-Unit scores on the Hunt task will indeed measure foreign students' command of English

structure effectively and validly. Another question is how effectively these scores can be estimated if multiple-choice approximations to the Hunt task are used.

Cloze procedure task. Variations of the Cloze procedure task, in which subjects are instructed to replace words deleted from prose passages, have long been of interest in language testing (see, for example, Taylor, 1953, 1956; Carroll, Carton, and Wilds, 1959; and Oller, 1973). Renewed interest in using the measure for foreign language testing was stimulated by Darnell (1970), when he developed a scoring procedure (Clozentropy) based on the frequency with which a large sample of American college students gave various substitutions for a specific omitted word. Strong interest in evaluating this type of measure for possible use in TOEFL has been expressed by some TOEFL Committee members and by some ETS staff working with the TOEFL program.

Questions regarding Cloze measures that guided the present study included the following: What are the advantages and disadvantages associated with different scoring methods--Clozentropy or Standard Cloze (accepting original word substitutions only)? Using written essay scores as criteria, how valid are the Cloze scores? Using both essay and Cloze scores as criteria, how valid are various multiple-choice approximations to the Cloze measures?

Because of various practical limitations, differences associated with the method of word deletion were not investigated. The procedure of deleting every tenth word from the Cloze passages was followed; alternative

procedures, such as the systematic deletion of nouns, adjectives, or function words, were not used. For the present study, an advantage of the nth-word deletion method was that it did not require the imposing of a priori constraints on the kinds of words to be tested. It provided, instead, a random-like sampling of words and of their immediate contexts.

Alternative Multiple-Choice Subtests

The development or selection of four alternative multiple-choice subtests that might fit into a future TOEFL was based on the questions and observations regarding content specifications that were noted earlier. These alternative subtests were administered as sections of an "Experimental TOEFL" and will be discussed in their order of appearance in that instrument.

Experimental Section X1, Sentence Comprehension. This section consists of test items taken from subsection T1a (Sentences) of Listening Comprehension sections of retired TOEFL forms. The only change was to present the questions or statements, as well as the answer choices, in the written mode.

This section served two purposes. One was to provide a partial check on whether the difficulty of Listening Comprehension items is indeed in the listening component of the task. If this is true, a test made up of equivalent items, but presented entirely in written form, should be much easier than the listening-based item statistics would suggest. The second purpose was to measure reading comprehension at the sentence level. If the sentence is the basic meaningful unit of connected prose, then a sentence comprehension measure may very closely approximate the

less wieldy Reading Comprehension measure described above. The questions concerning Experimental section X1 are directly related to these two purposes.

Experimental Section X2, Words in Context. Each item in this section consists of a complete sentence, with a target word or phrase underlined. The answer choices are alternative words or phrases for the underlined part of the sentence. For each sentence, the subject is instructed to ". . .find the one choice that will best replace the underlined part of the sentence, so that the basic meaning of the sentence remains the same."

This item format was used in part to meet the criticism that vocabulary items foster in foreign students an undue emphasis on vocabulary study of the kind required in preparing for a test of synonyms. By a test of vocabulary (words) in context, the emphasis is shifted to language study involving natural message units (sentences).

The format of Words in Context may have greater face validity than either of the formats in the TOEFL Vocabulary section (T3a, Sentence Completion, and T3b, Synonyms), because the task is like that often confronting any reader, in which he has both a word in question and its context to help him understand its meaning.

In writing the Words in Context items, an effort was made to use the kinds of words and contexts a student would be likely to encounter. This should further increase the face validity of the subtests and meet the criticism that vocabulary items too often test words the foreign student should not be expected to know.

Further comment regarding the difference between the Vocabulary-Sentence Completion and the Words in Context formats may be helpful. Although the two look much alike, and may or may not yield scores with

similar characteristics, they are logically quite distinct. The Words in Context item presents the candidate with a complete message unit. The distractors are words or sets of words which, when substituted for the underlined part of the sentence, yield a different message, but not necessarily an incorrect or anomalous sentence. The Vocabulary-Sentence Completion item on the other hand, presents the candidate with an incomplete sentence and, thus, an incomplete message unit. In order not to be "keyable," each distractor must yield an incorrect or anomalous result when it is used to complete the sentence.

Experimental Section X3, Combining Sentences. This task was developed to provide a multiple-choice sentence-embedding task that might approximate Hunt's rewriting task. The stem of each item consists of three to five short sentences. Each answer choice combines the short sentences in a different way. The subject's task is to ". . .choose the one long sentence that is the best combination of short sentences."

The main question in connection with the Combining Sentences task is, of course, how well scores on it estimate scores generated by the use of Hunt's rewriting task. Also of interest are how well the measure performs against essay writing criteria, and whether its pattern of correlations with other measures is logically satisfying.

Experimental Section X4, Paragraph Completion. This section consists of two multiple-choice variations of the Cloze task. Each task includes a reading passage with some of the words omitted and replaced by a numbered blank. On a facing page, a set of numbers corresponding to the numbered blanks is given, with each number followed by four words, one of which is the word originally fitting the numbered blank.

Questions concerning this item type are: How well do Paragraph Completion scores estimate Cloze scores? How well does the measure perform against essay writing criteria?

## Related Questions

To reduce the complexity of presentation, certain general questions have been held for this part of the discussion of the purpose and background of the study.

For virtually all of the questions that have been discussed in this section the answers may vary, depending on the language background of the candidates. Thus, a first general question may be asked: To what degree do the findings for candidates having an Indo-European first language hold for candidates from a non-Indo-European background?

A logical case has been presented for differentiated testing of English language skills. However, it is entirely possible that, for the great majority of candidates, the development of certain component skills may be so similar that there is no practical utility in having separate test sections for each. Thus, a second general question may be asked: How independent, in fact, are the component skills of English as a second language? This, in turn, has direct implications for a third general question: "How many separate scores should be reported in TOEFL?" The second and third general questions may be refined to be answered separately for subjects having different language backgrounds.

## PLAN AND PROCEDURE

The plan of the study called for the administration of a battery of measures including TOEFL, alternative multiple-choice and open-ended measures of English as a foreign language, and direct tests of speaking and writing performance in English. The tests, the subjects, and the procedures for administering and scoring the measures are described in this section of the report. The results, conclusions, and a discussion of the implications of the study are described in subsequent sections.

### Testing Materials

Each of the five TOEFL and four Experimental TOEFL subtests is an objective measure using a multiple-choice format. As a multiple-choice measure, each can be readily employed in a large-scale testing program such as TOEFL, but none can provide a direct estimation of a candidate's performance in the productive areas of English as a second language.

#### Multiple-Choice Measures Used in TOEFL

Section T1, Listening Comprehension. This section has three parts: Sentences, Dialogues, and Lecture.

There are two kinds of tasks in T1a, Sentences. One kind is answering a short question; the other is understanding a short statement. Each question or statement is presented in the spoken mode; the answer or paraphrase choices are given in written form.

Example I. When did Tom come here?

(A) By taxi.
(B) Yes, he did.
(C) To study history.
(D) Last night.

Sample Answer

I. [A] [B] [C] ■

Example II. John dropped the letter in the mailbox.

(A) John sent the letter.
(B) John opened the letter.
(C) John lost the letter.
(D) John destroyed the letter.

Sample Answer

II. ■ B C D

In T1b, Dialogues, the candidate hears a series of short conversations between two speakers. At the end of each conversation, a third voice asks a question about what has been said. The four possible answers to each question are given in written form.

Example III.

(man) Hello, Mary. This is Mr. Smith at the office. Is Bill feeling any better today?

(woman) Oh, yes, Mr. Smith. He's feeling much better now. But the doctor says he'll have to stay in bed until Monday.

(third voice) Where is Bill now?

(A) At the office.
(B) On his way to work.
(C) Home in bed.
(D) Away on vacation.

Sample Answer

III. A B ■ D

In T1c, Lecture, the candidate listens to a brief lecture, and is instructed to take notes as he might if he were attending a university lecture. A page is provided for his note-taking, at the top of which are written several names and terms that occurred in the lecture, of the kind a lecturer might write on the Chalkboard in class.

At the end of the lecture, the candidate opens his test book to a set of questions based on the lecture. He is allowed to use his notes while answering the questions.

T2, English Structure. In this section each problem consists of a short written conversation between two speakers, part of which has been omitted. Four words or phrases are given beneath the conversation, one of which will correctly complete it.

Example I. "John needs a pencil."
"He can use one -------."

Sample Answer

(A) of me
(B) my
(C) mine
(D) of mine

I. [A] [B] [C] [■]

Example II. "Did you remember Mary's birthday?"
"Yes, I -------."

(A) her sent a gift
(B) sent her a gift
(C) to her a gift sent
(D) a gift to her sent

II. [A] [■] [C] [D]

T3, Vocabulary. This section has two parts, Sentence Completion and Synonyms. Examples of T3a, Sentence Completion items, are the following:

Example I. A ------- is used to eat with.

Sample Answer

(A) plow
(B) fork
(C) hammer
(D) needle

I. [A] [■] [C] [D]

Example II. To escape is to get -------.

(A) away
(B) down
(C) up
(D) over

II. [■] [B] [C] [D]

Examples of T3b, Synonyms, are the following:

Example III. foolish

Sample Answer

(A) clever
(B) mild
(C) silly
(D) frank

III. [A] [B] [■] [D]

Example IV. a large branch of a tree

(A) straw
(B) limb
(C) bean
(D) vine

IV. [A] [■] [C] [D]

T4, Reading Comprehension. In this section, the candidate is given a series of paragraphs to read, each followed by several questions about what it means.

Sample paragraph. The White House, the official home of the President of the United States, was designed by the architect James Hoban, who is said to have been influenced by the design of a palace in Ireland. The building was begun in 1792 and was first occupied by President and Mrs. John Adams in November 1800. The house received its present name when it was painted white after being damaged by fire in 1814.

Example I. When was the White House first occupied?

(A) 1776
(B) 1792
(C) 1800
(D) 1814

Sample Answer

I. [A] [B] [■] [D]

Example II. According to the paragraph, the President's house was first painted white when

(A) President and Mrs. Adams requested that it be repainted
(B) it was repaired following a fire
(C) the architect suggested the new color
(D) it was remodeled to look like an Irish palace

II. [A] [■] [C] [D]

T5, Writing Ability. There are two parts to this section. Each problem in T5a, Error Recognition, consists of a sentence in which four words or phrases are underlined, and marked (A), (B), (C), or (D). The candidate is asked to identify the one underlined word or phrase that would not be acceptable in standard written English.

Example I. At first (A) the old woman seemed unwilling to accept (B) anything that (C) was offered her by my friends and I (D).

Sample Answer

I. [A] [B] [C] ■

Example II. After they had chose (A) the books they wished to read (B), the instructor told them (C) the principal points he wanted them to note (D).

Sample Answer

II. ■ [B] [C] [D]

In T5b, Sentence Completion, each problem consists of an incomplete sentence. Four words or phrases, marked (A), (B), (C), or (D), are given beneath the sentence. The candidate is to choose the word or phrase that best completes the sentence.

Example III. Because he had little education, his knowledge of the subject was -------.

(A) limited
(B) small in quantity
(C) minor
(D) not large at all

Sample Answer

III. ■ [B] [C] [D]

Example IV. At 7:00 tonight, a public lecture on nuclear physics will be delivered in the University auditorium by a --------.

(A) real informed man
(B) very authoritative guy
(C) prominent scientist
(D) person who knows a lot about it

IV. [A] [B] ■ [D]

Alternative Multiple-Choice Measures

The Experimental TOEFL subtests were developed at Educational Testing Service, specifically for the present study. A rationale for the inclusion of each experimental subtest was provided in the Purpose and Background section of this report. A description of the item format for each subtest follows.

Experimental Section X1, Sentence Comprehension. This subtest parallels Part A of the TOEFL Listening Comprehension section. However, the questions or statements, as well as the answer choices (options), are presented in written form. The examples shown above for the TOEFL Listening Comprehension section, T1a, apply as well to Experimental TOEFL, Section X1.

Experimental Section X2, Words in Context. Each sentence in this section has a word or phrase underlined. Four choices are given beneath the sentence. The candidate is to select the option that will best replace the underlined part of the sentence, so that the basic meaning of the sentence remains the same.

Example I. He discovered a new route through the mountains.

(A) wanted
(B) found
(C) traveled
(D) captured

Sample Answer

| | A | B | C | D |
|---|---|---|---|---|
| I. | | ▮ | | |

Example II. Their success came about as a result of your assistance.

(A) according to
(B) before
(C) because of
(D) during

| | A | B | C | D |
|---|---|---|---|---|
| II. | | | ▮ | |

Experimental Section X3, Combining Sentences. Each item in this section consists of a group of short, related sentences. Four long sentences are given below the group of short sentences. For every item, each wrong option presents a message which differs from that conveyed by the short sentences in the stem although it does not necessarily depart from standard usage. The candidate's task is to choose the option that is the best combination of the short sentences.

Example I. John is in the store. It is a hardware store. Fred is also in the store. They are buying tools.

(A) John is buying tools from Fred in the hardware store.
(B) John is buying hardware tools from Fred in the store.
(C) John and Fred are buying tools in the hardware store.
(D) John and Fred are buying hardware tools in the store.

Sample Answer

| | A | B | C | D |
|---|---|---|---|---|
| I. | | | ▌ | |

Example II. There was an accident. A car went off the road. A young man drove it. The car belonged to his father.

A young man . . .

(A) accidentally drove his father's car off the road.
(B) accidentally drove the car off his father's road.
(C) drove his father's accidental car off the road.
(D) drove the car off his father's accidental road.

Sample Answer

| | A | B | C | D |
|---|---|---|---|---|
| II. | ▌ | | | |

Experimental Section X4, Paragraph Completion. This section is made up of two reading passages, each with some words omitted and replaced by a numbered blank. On a facing page, a set of numbers corresponding to the numbered blanks is given, with each number followed by four words. For each numbered blank, the candidate is to choose the word that best fits the context.

Examples I and II. For good reason, historians use the I of writing to mark the divide between history II prehistory.

| I. | | II. | |
|---|---|---|---|
| | (A) job | | (A) in |
| | (B) effort | | (B) or |
| | (C) decision | | (C) and |
| | (D) invention | | (D) from |

Sample Answer

| | A | B | C | D |
|---|---|---|---|---|
| I. | | | | ■ |
| II. | | | ■ | |

Open-Ended Objective Measures

It is sometimes assumed, incorrectly, that objective measures are necessarily cast in the multiple-choice item format. Two interesting exceptions to such a proposition are the Hunt rewriting task and the Cloze procedure. Both tasks impose more constraint than a free-response essay assignment would, but the responses called for are, nevertheless, open-ended rather than multiple-choice, and they are objectively scorable. As measures of writing ability, the tasks are more direct than multiple-choice measures, but less direct than an essay-writing assignment. As might be expected, they are also intermediate with respect to scoring costs.

Hunt's Aluminum passage. With the permission of Kellogg Hunt, the Aluminum passage reported in his study (1970b) was used in the present study. The same directions, translated into Spanish and Japanese for the two subject groups, were also used. The directions in English read as follows:

> Directions: Read the passage all the way through. You will notice that the sentences are short and choppy. Study the passage, and then rewrite it in a better way. You may combine sentences, change the order of words, and omit words that are repeated too many times. But try not to leave out any of the information.

The passage presented to the subjects consisted of 32 very short sentences of connected discourse. The first portion is as follows:

> Aluminum is a metal. It is abundant. It has many uses. It comes from bauxite. Bauxite is an ore. Bauxite looks like clay.

Subjects combined these sentences with varying amounts of embedding, and with varying degrees of success in retaining the original units of information. All six of the above sentences could be embedded into a single sentence, yielding something like the following: Aluminum, an abundant metal with many uses, comes from bauxite, a clay-like ore.

Hunt's principal measure of "syntactic maturity," Words per T-Unit, was also adopted for use in the present study. Hunt (1970b) defines the T-unit, or "minimal terminable unit," as ". . . one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it . . . . So cutting a passage into T-units will be cutting it into the shortest units which it is grammatically allowable to punctuate as sentences [p. 4]." Largely an objective measure, the Words per T-Unit measure requires some initial judgment, but once T-unit boundaries have been agreed upon, it amounts basically to a word count. Furthermore, very high agreement on assigned T-unit boundaries can be reached by trained readers judging the rewritten passages.

Prior to the actual scoring of Words per T-Unit, Hunt's procedure deleted extraneous, unintelligible, or inaccurate passages. When these were found, the entire sentence involved was deleted. In the present study extraneous sentences were also eliminated, but faulty sentences were retained.

An additional score, "High K's," was generated for this study, to indicate the number of short kernel sentences (K's) that were adequately represented in each subject's rewriting of the Aluminum passage. In contrast to the Words per T-Unit score, the High K's score is definitely subjective and as such requires a considerable amount of time and effort to reach satisfactory levels of agreement in judgment.

Cloze passages. Two prose passages, each about 280 words in length were used as the basis for the Cloze procedure tasks. The first sentence in each passage was left intact to provide an adequate context to begin the task. In the remainder of the passage every tenth word was replaced by a blank, yielding 25 blanks for each passage.

The original instructions for the Cloze tasks are shown below. These were translated into Spenish or Japanese with the exception of the italicized words and the handwritten answers, which were retained in English.

> Instructions: This section contains two reading passages with some words omitted. Fill in the one word that you think best completes each blank in the passage below.
>
> EXAMPLE
>
> *John came to school _____ bus. He thought it was the _____ way.*
>
> You would probably write *"by"* in the first blank. You might write *"safest"* or *"quickest"* or *"cheapest"* in the second blank. Use any word which seems to be correct to you.
>
> Thus, you might complete the two sentences:
>
> *John came to school by bus. He thought it was the safest way.*

We suggest that you scan an entire passage, go over it again filling in the easy blanks, then go back a third time filling in the difficult ones.

Subjective Measures

The English language productive skills, speaking and writing, are fundamental to the academic success of the foreign student. The language proficiency interview and the essay writing tasks developed for this study required actual performance in these skills. These subjective measures were designed to provide direct and reliable, as well as feasible, estimates of the speaking and writing abilities needed in the academic setting. Although they were somewhat removed from the speaking and writing tasks encountered in the classroom, the measures had the advantages of assigning the same tasks to all subjects and of providing scores not influenced by the extraneous factors influencing a criterion such as teacher-assigned grades.

Interviews. Performance in English language conversation was measured by means of structured, tape-recorded interviews which were conducted at the testing sites. The performance judgments were subsequently made by staff members at Educational Testing Service. The judgment scales were Accent, Grammar, Vocabulary, Fluency, and Overall Communication. Scores for the latter three scales were derived separately for the general or narrative part of the interview and for the academic or technical part.

The procedure for conducting and evaluating the interviews was adapted from the Peace Corps Language Proficiency Interview, which in turn derives in part from the Absolute Language Proficiency Rating

prepared by the Foreign Service Institute for the classification of officers of the United States Department of State (Rice, 1959; Wilds, undated).

The interviews were typically 20 to 30 minutes in length. In Peru and Chile, they included four stages. The first was an exploratory stage of about 2 to 4 minutes, designed to put the interviewee at ease. Second, there was a narration task of about 5 minutes duration, in which the candidate was asked to relate in English a story from a Spanish-language comic book. (He was allowed to select one of eight such stories beforehand and was given 10 minutes preparation time immediately before the interview. He could look at the comic book while telling the story, but was advised to use the pictures only, because an attempt to translate rather than narrate would be likely to hinder, rather than facilitate, his narration.) Third was a 4 to 6 minute paraphrasing task in which a graded series of increasingly complex sentences was read to the candidate. After each sentence was read, the subject's task was to paraphrase it. The fourth was an analytical stage, 6 to 10 minutes long. In order to provide an indication of the student's English language skills that would be appropriate in an academic setting, the conversation was directed toward the student's plans and major areas of study, the ideas and writers of interest in his or her field, etc. The second and third parts of this interview are innovations not found in the FSI and Peace Corps interviews.

Interviews in Japan followed the same pattern, except that the second stage was a period of general conversation, rather than the narrative task.

For the narrative or general part and for the analytic part of the interview, interviewers were instructed to try to get a good representative sample of the candidate's capacity in specific language areas. Thus, the interview was to give the student the opportunity to show his knowledge of verb forms, person-subject-object agreement, the formation and use of adjectives and adverbs, and other aspects of grammar. Similarly, it was to be conducted so that it would provide a good sample of the candidate's general vocabulary and academic or technical vocabulary. Interviewers were informed that the candidate would also be judged on accent, fluency, and overall communication, but that interviewers did not need to try in any way to help the student demonstrate his performance in these areas.

Details regarding the administration and subsequent scoring of the tape-recorded interviews will be given below.

Essays. Four essay tasks were assigned, with 10 minutes provided for each.[1] The assignment of several short essays rather than one or two longer ones was based on evidence (e.g., Godshalk, Swineford, and Coffman, 1966) that essay ratings vary significantly from topic to topic. Two of the essays used pictures for stimuli; the other two did not.

Instructions written in Spanish or Japanese were provided for each of these tasks.

---

[1] These tasks, based on materials developed by John Carroll and an international committee, are described in the document, International Association for the Evaluation of Educational Achievement, Phase II, Stage 3, French as a Foreign Language, June 1970. Permission to use these materials was received from the I.E.A. Bureau, communicated through T. Neville Postlethwaite, Executive Director, I.E.A. Wenner-Gren Center, Stockholm, Sweden.

In the first essay task, the subject was presented a sequence of three pictures in comic-book format and was instructed to describe (in English) what is happening in the set of pictures. The second task called for writing a dialogue between two boys and incorporating certain words or expressions listed in the directions, such as "beautiful day," "to take a walk," and "bicycle." The third presented a single picture, in which a boy's bicycle has just been damaged by an automobile. The candidate was instructed to describe what he thought led up to the event, what is happening in the picture, and what will happen next. The fourth exercise called for a composition comparing the advantages of living in the country and in a large city. The candidate was provided certain terms, such as "peacefulness" and "department stores," to be included in the essay.

## Subjects

Participants in the study included 98 Peruvians, 145 Chileans, and 199 Japanese, to whom Form TEF4 of TOEFL was administered in Lima, Santiago, and Tokyo, respectively.

Several considerations led to the inclusion of subjects from the two language backgrounds, Spanish and Japanese. If two very different first languages are represented, one Indo-European and the other not, the findings can be more generally interpreted than if a single language or only closely related languages are represented. Limiting the backgrounds to two made it possible to have enough subjects from each language group to allow meaningful and useful analyses of complex questions. The Spanish and Japanese backgrounds were chosen because they meet the criterion of being distinctly different from one another, and because a high volume

of TOEFL candidates comes from each. The latter fact made it possible for data to be gathered in a few central locations from a sufficiently large number of candidates. It also meant that the findings specific to the respective language backgrounds would be directly relevant to a large number of people.

The decision to administer TOEFL and the experimental research measure to candidates in their own countries was based on the following considerations. First, the students' performance in English as a second language would not have been influenced by varying amounts of formal and informal exposure to English in the United States. Second, it was much easier to test the required number of subjects in a few locations. Finally, testing overseas avoided the problem of restriction of score range inherent in testing foreign students already accepted for study on an American campus, whose admission was usually based in part on TOEFL scores.

An Individual Background Questionnaire was developed and translated into Spanish and Japanese. All participants completed the questionnaire, which asked their age, sex, parents' education, major area of study, level of education completed, recent school grades, and the source and amount of formal English language instruction and informal exposure to English. They were also asked to estimate their general level of competence in reading, writing, listening, and speaking English. The questionnaire responses are discussed in the Results and Conclusions section of this report.

## Administration of Measures

The recruitment of subjects and the administration of all measures except TOEFL were carried out by collaborators and their support staffs at the Instituto Peruano de Fomento Educativo (IPFE) in Lima, Peru, at the Universidad de Chile in Santiago, Chile, and at Language Education Associates, Incorporated, and the International Education Center, both in Tokyo, Japan.

Recruiting was carried out primarily by contacting students who had applied to take the October 1971 administration of TOEFL and who would take that test in one of the cities noted above.

Instrumentation, as used in the study, was grouped into the following units: (1) TOEFL, Form TEF4, with the five sections described earlier in this part of the report. Total testing time, 2 hours, 20 minutes. (2) An Experimental Test of English as a Foreign Language, with the four sections also described earlier. Testing time, 2 hours, 25 minutes. (3) An Experimental Test of English Writing Ability. (There were Spanish and Japanese versions of this instrument, with directions in the appropriate language.) This test contained the Hunt rewriting task, the Cloze procedure passages, and the four essay assignments. Testing time, 2 hours. (4) Individual Background Questionnaire, Spanish and Japanese versions. Completion time, about 15 minutes. (5) A structured English proficiency interview. Interview time, 20 to 30 minutes.

In Peru, all participants in the study took TOEFL at the regular October 1971 administration of that examination. In Chile and Japan, about half of the participants took TOEFL at the regular administration, and half took it at a special administration under standard conditions. All remaining instruments were administered at the institutions noted above. The written tests were given within three days after TOEFL was administered. These were followed by the interviews and background questionnaires, most of which were completed within three weeks.

All interviewers were native speakers of English. In Peru and Japan, most of the interviewers were recruited from American embassy personnel or members of their families. They were trained by persons experienced in administering the Peace Corps Language Proficiency Interview. Each began with practice interviews, following the procedures developed for the present study. In Chile, the interviews were conducted by professors Ewer and Hughes-Davies and three of their staff members. All were experienced in the use of interviews to assess the English language skills of Chilean students.

## Scoring Procedures

### Multiple-choice Measures

The five TOEFL and four Experimental TOEFL measures were scored according to the standard procedures applied to multiple-choice tests. Because "rights only" scoring is used for TOEFL, the same procedure was applied to the Experimental TOEFL. (The fact that there is no penalty for guessing was indicated in the candidates' instructions for both examinations.)

Open-Ended Objective Measures

Hunt's Aluminum passage. Score sheets and instructions used to obtain the Words per T-Unit and High K's values are shown in Appendix A. For each candidate's rewriting or protocol, the Words per T-Unit value was obtained by first striking out extraneous sentences, then marking the T-unit boundaries, then counting the number of words and T-units and computing the ratio of Words/T. The adequacy with which the information in each kernel sentence or K was expressed in a given protocol was judged as high, medium, low, or absent, using the instructions shown in Appendix A. The High K's score used in subsequent analyses was the total number of the 32 K's judged to have been effectively expressed in the protocol. The High K rating was determined by whether the K in question was clearly and unambiguously stated; it did not necessarily have to be stated in standard English.

The Hunt protocols obtained for the study were scored for Words per T-Unit and for High K's by two scorers, who first practiced with protocols obtained through pretesting. Approximately one-third of the protocols for the study were rated independently by both scorers. This provided a basis for estimating scorer reliability, and for the scorers and the author to meet periodically to consider differences in scoring and subsequently to refine scoring procedures. The other two-thirds of the protocols received only a single rating.

Protocols were randomly assigned to batches for scoring, and were randomly ordered within each batch. Those in batches scored by both scorers were assigned two random orderings, one for each judge. These randomization procedures were used to prevent any systematic order effect, such as might occur through shifting standards over a period of making judgments.

Cloze passages. As a preliminary step for Clozentropy scoring, the Cloze passages were administered to 260 American college students. Their responses to each blank were tallied, and a "dictionary" was made up for that blank, listing the response words and the frequency with which they occurred. For example, one of the blanks appeared in the second Cloze passage, as follows:

*But, as _____ morning advanced, the strength of the gusts diminished.*

The dictionary for this blank, based on American students' responses, was:

| Word | Frequency | Word | Frequency | Word | Frequency |
|---|---|---|---|---|---|
| each | 1 | late | 3 | Soon as | 1 |
| early | 21 | mid | 1 | the | 230 |
| first | 1 | new | 1 | usual | 1 |

Using Reilly's (1971) simplification of Darnell's (1970) Clozentropy scoring procedure, a foreign student's response to a given blank was scored as follows. If he did not fill in the blank, or if his response was a word not listed in the dictionary for that blank, he was given a score of zero. If his response was one of the words in the dictionary, his score for that blank was the logarithm of the frequency associated with his response word.

In the Standard Cloze scoring procedure, the foreign student's response was compared with the word that had been omitted. If his response matched the original word for a given blank, the student was given a score of 1; if it did not match, his score was zero.

Both scoring procedures were carried out by computer, and were not

unreasonably expensive. The expensive and time-consuming part of obtaining scores for Cloze responses lay in the work of preparing them for keypunching, and in the keypunching itself.

Subjective Measures

Interviews. Each candidate's proficiency in spoken English was rated independently by three listeners who used the interview rating sheet and the proficiency descriptions shown in Parts 1 and 2 of Appendix B.

In preparing each interview tape for the listeners, the author or a research assistant listened to the complete tape, and then marked two segments to be used in assigning the proficiency ratings. One 4 or 5 minute segment was selected from what appeared to be the candidate's best performance on the narrative or general conversation part of the interview, and the other from the best portion of the academic or technical part. This procedure reduced by two-thirds the listening time needed to assign ratings, while ensuring that the three sets of ratings for each candidate were still based on the same language samples.

There were ten listeners, including the author. He and two research assistants developed eight training tapes and eight practice tapes for the initial training of the other seven listeners. Each training tape was accompanied by an information sheet which provided the scores agreed upon by the three ETS staff members and comments illustrating what might be written in the lower half of the rating sheet regarding the bases for the scores assigned. Four of these information sheets are shown in Appendix B. After working with the training tapes, all listeners rated the practice tapes independently. Group training sessions were held, in which differences in rating were discussed and resolved.

The proficiency descriptions were designed to keep the interview ratings on an absolute, criterion-referenced scale of English language proficiency. In order to reduce the tendency to move to a normative scale, or for the rating of a given interview to be unduly influenced by the proficiency demonstrated in immediately preceding tapes, raters regularly referred to these descriptions when scoring the interviews.

It was found that the listeners were more comfortable about assigning ratings if between-level values could be used. The scale was therefore expanded to include 16 levels: 1, 1+, 2-, 2, 2+, . . . , 5+, 6-, and 6. The six unmodified numbers are still anchored by the proficiency descriptions.

In making their ratings, listeners were urged to keep the several scales independent. It was pointed out, however, that the Communication scale, with its emphasis on the ability to convey meaning or content effectively, was bound to be influenced in varying degrees by the other proficiencies that the listeners were rating.

For scoring purposes, the taped interviews were grouped into 20 batches of 20 tapes each, and a final batch of 22. In assigning tapes to batches, proportional representation of the three nationalities was imposed, so that each of the 20 batches had four or five tapes from Peru, six or seven from Chile, and eight or nine from Japan. Then, for each batch, the designated number of tapes for each nationality was randomly selected from the set of tapes from the respective countries.

Tapes were assigned to listeners in the batches of 20. Each batch was listened to independently by three listeners. Whenever a batch was assigned to a new listener, a new random list of numbers 1-20 was also

assigned, indicating the order in which the tapes were to be rated. The final batch of tapes was used for continued training and to keep all scoring coordinated. There were five coordinating sessions in which all raters met and listened collectively to three to five tapes. Each rated the tapes independently. After each tape was played and rated, the ratings were tallied, the differences discussed, and new rating problems and considerations were examined.

Essays. As noted above, the candidates wrote four short essays, each on a different topic. Each essay was rated independently by two or more readers on each of the following scales:

E1a Content (Quantity). The number of ideas and concepts expressed, the degree of elaboration, etc.

E1b Content (Quality). The adequacy and interest value of the story line, internal consistency, how well the story "comes across," and whether it is easy to follow.

E2a Form (Quantity). The range of vocabulary and of grammatical structures used.

E2b Form (Quality). The appropriateness and effectiveness of the vocabulary and structures used.

The essays were scored by a procedure using several readers for each paper, each making rapid, holistic judgments regarding one of the four scales described above. The procedure was derived from that developed for the Godshalk et al. study, The Measurement of Writing Ability (1966), and since refined in essay readings conducted as part of the College Board's English Composition Test.

The 21 participating readers were all experienced with essay-grading procedures like those used in this study. All had participated in other ETS

essay readings, in cooperation with the ETS Essay Reading Group.

Readers were organized into four groups or tables, each with a table leader. Readers at one table read only for scale Ela, Content (Quantity), those at another table read only for scale Elb, Content (Quality), and so on. By this arrangement, readers at a given table could develop a "set" for the kind of judgment they were to make, keeping it independent of the other three kinds of judgment called for.

Prior to the readings, the table leaders were oriented by the chief reader, and each prepared in turn to orient the readers at his or her table. A 6-point normative-reference scale was introduced, in which level 2 represented the typical or median paper for the lower half of the essays, and level 5 the typical essay among the upper half. The score for a lower-half essay could be "shaded down" to level 1, or "shaded up" to level 3. Similarly, an upper-half essay could be graded 5, or shaded to either 4 or 6. After being assigned to one of the four scales, each table leader scanned a sample of Topic I essays, then selected essays that epitomized each of the six levels of writing proficiency. These were subsequently used to orient the table readers, and to serve as reference points for the six levels. The same procedure was followed in preparing to read and score essays on Topics II, III, and IV.

All essays on Topic I were scored first, then those on Topic II, and so on. While essays on a given topic were being scored, they were moved in small batches from one table to another until each had been scored independently by two readers at each table. As essays went through each step of this sequence of eight readings, the order was scrambled to prevent any order effect, and each successive score was covered up so that

the following ratings would not be influenced by those preceding.

Whenever the two ratings for a given essay on a given scale were separated by two or more intervening categories (e.g., if the scores were 1 and 4, or 2 and 5), a third independent rating was obtained. In those cases, the third rating and the earlier one closest to it were retained, and the other score dropped.

## RESULTS AND CONCLUSIONS

To provide a background for discussing test results, the subject groups will be described first, primarily from their individual responses to the background questionnaire. Then summary test statistics from the several multiple-choice measures, open-ended objective measures, and subjective measures will be examined and compared, as well as the inter-relationships among them. Finally, conclusions regarding each of the measures will be discussed in turn.

As noted earlier, participants in the study included 98 Peruvians, 145 Chileans, and 199 Japanese, all of whom were administered form TEF4 of TOEFL. Because of some attrition, respondents to the background questionnaire numbered 86, 136, and 196, respectively. The smallest numbers of candidates responding to test materials in addition to TOEFL TEF4 were 91 for Peru, 140 for Chile, and 187 for Japan. The N's for the separate measures are given in the appropriate tables. In the data analyses, missing data procedures were used where needed.

The substantial linguistic differences between the Latin American and Japanese participants provide a basis for partitioning between findings that are unique to those having either Spanish or Japanese language backgrounds and those that hold for both languages and that may therefore be more readily generalized to subjects having yet other language backgrounds. Throughout the following discussion similarities and contrasts between Spanish and Japanese data will be pointed out, as well as those between

data from the two Spanish language groups, Peruvians and Chileans. Although these observations will be of interest for what they indicate about the English language skills of the respective groups, they will be focused primarily on implications regarding the present and alternative TOEFL measures under investigation.

## Background Questionnaire Responses

Responses to general background questions are reported by language group in Table 1. In discussing these data, the differences between the Peruvian and Chilean samples will be noted where appropriate.

The responses to question 3 show that less than one percent of the sample were from families where English was the language used in the home. Among Peruvians, the non-Spanish languages indicated were Chinese and Quechua (Inca). Among Chileans, the non-Spanish language used in the home was usually either German or Italian. Among the Japanese subjects, the non-Japanese language used in the home was Chinese in four instances, and Tagalog in the remaining one.

For the two language groups, 77 percent of the Spanish and 61 percent of the Japanese indicated having completed three or more years of higher education. Slightly over three-fourths of the Spanish subjects planned to earn a doctorate, but this was true for only one-tenth of the Japanese. This difference apparently reflects the difference in use of the doctoral degree in the respective countries.

Regarding college majors, the natural sciences were favored by the Spanish subjects, but not by the Japanese. Engineering as a major field was also considerably more popular in the Spanish group than in the Japanese, where higher percentages are shown in the humanities and social sciences.

Table 1

Responses to General Background Questions, by Percentages

| Background question | Language group Sp | Jpn |
|---|---|---|
| 1. Age | | |
| less than 21 | 15 | 18 |
| 21-23 | 31 | 40 |
| 24-26 | 22 | 19 |
| 27 or older | 23 | 22 |
| No response | 9 | 0 |
| 2. Sex | | |
| Male | 70 | 60 |
| Female | 30 | 40 |
| 3. Language in the home | | |
| Spanish | 92.3 | 0.0 |
| Japanese | 0.0 | 97.0 |
| English | 0.5 | 0.0 |
| Chinese | 1.4 | 2.0 |
| German | 2.3 | 0.0 |
| Italian | 1.4 | 0.0 |
| Other | 2.8 | 0.5 |
| 4. Education completed | | |
| Secondary school or less | 12 | 12 |
| 1-2 years higher education | 8 | 26 |
| 3-4 years higher education | 24 | 51 |
| 1 or more years graduate school | 53 | 10 |
| No response | 4 | 1 |

| Background question | Language group Sp | Jpn |
|---|---|---|
| 5. Educational objective | | |
| Bachelors or less | 4 | 39 |
| Licentiate/ Masters | 14 | 29 |
| Doctorate | 77 | 10 |
| Other | 0 | 6 |
| No response | 6 | 16 |
| 6. Secondary school major | | |
| College prep | 56 | 57 |
| Community/ vocational | 3 | 6 |
| General | 29 | 36 |
| Other | 1 | 1 |
| No response | 11 | 1 |
| 7. Higher education major | | |
| Agriculture | 6 | 1 |
| Economics | 10 | 18 |
| Education | 10 | 2 |
| Engineering | 29 | 16 |
| Humanities | 1 | 2 |
| Language/Lit. | 5 | 24 |
| Natural Science | 12 | 4 |
| Social Science | 6 | 12 |
| Professional Services | 8 | 7 |
| No response | 12 | 15 |

Note: Tabulated responses are from questionnaires administered to 86 Peruvian, 136 Chilean, and 196 Japanese participants.

Self-reported grades in English as a second language and self-ratings on English language skills are given in Tables 2 and 3. The self-reported grades in English reading and writing were quite similar between the two language groups, but the Japanese tended to rate themselves somewhat lower than did the Spanish in English listening and speaking. In their self-ratings on all four English language skills (newspaper reading, essay writing, listening, and conversing), the Japanese tended to rate themselves much lower than did the Latin American candidates.

Table 2

Self-Reported Grades in English as a Foreign Language

| Level of grades in EFL | Reading | | Writing | | Listening | | Speaking | |
|---|---|---|---|---|---|---|---|---|
| | Sp | Jpn | Sp | Jpn | Sp | Jpn | Sp | Jpn |
| Excellent | 32 | 35 | 30 | 30 | 29 | 14 | 28 | 18 |
| Good | 42 | 35 | 38 | 32 | 34 | 29 | 28 | 31 |
| Fair | 16 | 17 | 20 | 22 | 19 | 21 | 27 | 18 |
| No response | 8 | 13 | 11 | 16 | 17 | 36 | 17 | 33 |

Note: N = 222 Spanish-language (86 Peruvian and 136 Chilean), and 196 Japanese-language participants.

Although the questionnaire data have not been separated into Peruvian and Chilean responses, taped interviews and other sources of information indicate that the two groups are distinct in certain respects. Participants in the Chilean sample account for most of the Spanish subjects already in graduate school and aspiring to a doctorate. They were typically middle-class or higher, and urban (from Santiago or its suburbs). By

contrast, the Peruvian subjects were typically from smaller towns or cities some distance from Lima, and were lower-middle class or below.

Table 3

Self-Ratings on English Language Skills

| Level of difficulty in stated activity | Newspaper reading | | Essay writing | | Listening | | Conversing | |
|---|---|---|---|---|---|---|---|---|
| | Sp | Jpn | Sp | Jpn | Sp | Jpn | Sp | Jpn |
| Easily | 40 | 15 | 44 | 5 | 29 | 10 | 29 | 5 |
| With some difficulty | 56 | 82 | 47 | 79 | 57 | 83 | 57 | 77 |
| With much difficulty | 2 | 3 | 8 | 16 | 12 | 6 | 13 | 17 |

Note: N = 222 Spanish-language and 196 Japanese-language participants.

## Summary Test Statistics

Summary test statistics will be presented for the three subject groups for the multiple-choice measures found in TEF4 and in the "Experimental TOEFL," the open-ended objective measures derived from Hunt's aluminum passage and the two Cloze passages, and for the subjective interview and essay measures.

### Multiple-Choice Measures

Means and standard deviations of multiple-choice scores for the three subject groups are shown in Table 4. "World" data are also given for the five major TOEFL scores, T1 through T5. The latter are from a spaced sample of 1000 candidates who took the October 1971 TOEFL examination, Form TEF4, as reported in ETS Statistical Report 71-112 (Swineford, 1971).

Means and Standard Deviations of Multiple-Choice Measures

| Multiple-choice measure | Items | Min.[a] | Mean | | | | Standard deviation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Peru | Chile | Japan | World | Peru | Chile | Japan | World |
| **Regular TOEFL sections** | | | | | | | | | | |
| T1 Listening Comprehension | 50 | 40 | 22.9 | 28.7 | 30.9 | 30.1 | 11.7 | 10.5 | 7.5 | 8.8 |
| T2 English Structure | 40 | 20 | 16.1 | 23.0 | 24.4 | 23.8 | 9.1 | 8.2 | 5.6 | 7.0 |
| T3 Vocabulary | 40 | 15 | 19.4 | 24.1 | 20.3 | 22.4 | 7.3 | 5.8 | 6.5 | 6.9 |
| T4 Reading Comprehension | 30 | 40 | 14.0 | 19.2 | 17.6 | 16.5 | 6.3 | 4.8 | 4.2 | 5.0 |
| T5 Writing Ability | 40 | 25 | 13.8 | 19.7 | 21.0 | 21.6 | 7.6 | 7.4 | 5.7 | 6.7 |
| (N) | | | (98) | (145) | (199) | (1000) | | | | |
| **Regular TOEFL subsections** | | | | | | | | | | |
| Listening Comprehension | | | | | | | | | | |
| T1a Sentences | 20 | 10 | 8.3 | 11.1 | 12.1 | | 5.6 | 4.7 | 3.4 | |
| T1b Dialogues | 15 | 12 | 7.0 | 8.5 | 9.2 | | 3.7 | 3.7 | 2.6 | |
| T1c Lecture | 15 | 18 | 7.6 | 9.1 | 9.6 | | 3.3 | 3.1 | 2.8 | |
| Vocabulary | | | | | | | | | | |
| T3a Sentence Completion | 15 | | 7.3 | 9.1 | 7.2 | | 2.9 | 2.6 | 2.7 | |
| T3b Synonyms | 25 | | 12.0 | 15.0 | 13.1 | | 4.8 | 4.0 | 4.4 | |
| Writing Ability | | | | | | | | | | |
| T5a Error Recognition | 25 | | 8.6 | 12.2 | 14.1 | | 4.8 | 4.6 | 3.7 | |
| T5b Sentence Completion | 15 | | 5.2 | 7.5 | 6.9 | | 3.5 | 3.8 | 3.1 | |
| **Experimental TOEFL** | | | | | | | | | | |
| X1 Sentence Comprehension | 30 | 25 | 21.7 | 27.1 | 26.6 | | 7.0 | 3.2 | 2.9 | |
| X2 Words in Context | 30 | 30 | 18.5 | 24.5 | 22.2 | | 6.6 | 4.2 | 4.2 | |
| X3 Combining Sentences | 30 | 40 | 18.3 | 23.0 | 22.1 | | 6.2 | 4.1 | 3.3 | |
| X4 Paragraph Completion | 50 | 50 | 24.5 | 32.5 | 28.9 | | 9.4 | 6.8 | 7.4 | |
| (N) | | | (91) | (145) | (197) | | | | | |

Note: Mean total scores on TEF4 were 411, 483, 480, and 479 for Peru, Chile, Japan, and the "World," respectively.

[a]Vocabulary and Writing Ability subsections were not separately timed.

[b]"World" data are from a spaced sample of 1000 of the 14,134 candidates who took the October 1971 TOEFL.

The rather consistent differences observed between subject groups in standard deviation values systematically influenced the intercorrelations among the measures discussed below. Peruvian standard deviations were largest for all five TOEFL subtests and for the four Experimental TOEFL subtests. Chilean standard deviations were larger than those of the Japanese for four of the five regular TOEFL measures, but for only half of the Experimental TOEFL measures.

The total TOEFL score means were nearly identical for the Chilean, Japanese, and World samples, with the Peruvian mean about 70 points lower. These relative magnitudes generally approximated those averaged over several years for the respective geographic areas as reported in the 1973 Manual for TOEFL Score Recipients. Given the difference in total TOEFL score, it is not surprising that the Peruvian means were lowest on all nine objective measures. When comparing Chilean and Japanese means on TOEFL, it is interesting to note that the Japanese were slightly higher on Listening Comprehension, English Structure, and Writing Ability, but scored about one-half standard deviation below Chileans on Vocabulary. When we note further that on Vocabulary the Peruvian mean was nearly equal to that of the Japanese, it would seem that a vocabulary test puts the Japanese at a disadvantage when compared to Spanish-background subjects. This observation raises the question of whether such differences in test scores associated with the candidates' first language should be considered as evidence of possible unfairness of specific subtests in a test of English as a foreign language. What if, for example, advantages associated with cognates and similarities in idiom allowed subjects from Indo-European language backgrounds to score higher on vocabulary than subjects from

non-Indo-European language backgrounds, although they did not score higher in certain other language areas such as "English Structure"? It would seem reasonable to suggest that such differences do not necessarily imply unfairness. Ideally, a test of English as a foreign language will assess the foreign student's English language skills needed for study at American campuses, regardless of how difficult or how easy it is for him to acquire these skills.

Mean score differences observed for the TOEFL Listening and Vocabulary measures were also shown for their respective subtests. However, the better showing of the Japanese in Writing Ability appeared only in the Error Recognition (T5a) component.

Experimental Section X1, Sentence Comprehension, parallels TOEFL subsection T1a, Listening Comprehension/Sentences, both in content and difficulty level. However, the stimuli in T1a were presented in the spoken mode, whereas parallel stimuli in experimental section X1 were presented in written form. In going from the spoken to the written mode of presentation, the Peruvian, Chilean, and Japanese mean scores rose from 42, 55, and 60 percent of the total number of items, to 72, 90, and 89 percent, respectively. These marked shifts suggest that much of the difficulty in answering TOEFL Listening Comprehension/Sentences items was indeed due to the listening component of the task.

Reliability indices for the multiple-choice measures are shown in Table 5. The reliabilities were generally high on all nine measures for each subject group. In all instances the Peruvian reliabilities were highest; the Chilean indices usually fell between the Peruvian and

Table 5

Reliability Indices[a] of Multiple-Choice Measures

| Multiple-choice measure | | Subject group | | | |
|---|---|---|---|---|---|
| | | Peru | Chile | Japan | World |
| | Regular TOEFL | | | | |
| T1 | Listening Comprehension | .94 | .93 | .85 | .88 |
| T2 | English Structure | .92 | .89 | .78 | .83 |
| T3 | Vocabulary | .88 | .81 | .82 | .83 |
| T4 | Reading Comprehension | .86 | .79 | .69 | .79 |
| T5 | Writing Ability | .89 | .86 | .76 | .82 |
| | (N) | (98) | (145) | (199) | (1000) |
| | Experimental TOEFL | | | | |
| X1 | Sentence Comprehension | .91 | .79 | .71 | |
| X2 | Words in Context | .90 | .81 | .76 | |
| X3 | Combining Sentences | .88 | .77 | .66 | |
| X4 | Paragraph Completion | .90 | .83 | .83 | |
| | (N) | (91) | (145) | (197) | |

[a]Computed by Kuder-Richardson formula 20.

Japanese in magnitude. These between-country differences would be expected as a result of the restriction-of-range indicated by the differences in standard deviation for observed scores among the respective subject groups.

Open-Ended Objective Measures

Summary statistics for the Hunt aluminum passage scores (High K's and Words per T-Unit) and for the Cloze passage scores (Clozentropy and Standard Cloze) are given in Table 6. Mean scores for each of the measures derived from the Hunt aluminum passage task were nearly the same for Chile as for Japan, but those for Peru were again lower. The mean Words per T-Unit scores for Peru, Chile, and Japan of 9.7, 11.0, and 10.9 may be compared to those observed by Hunt (1970b) for American children in grades 4, 8, and 12 of 8.6, 11.5, and 14.4.

For the measures derived from interview and essay judgments and from the Cloze passages, the standard deviations showed a marked rank-ordering, with Peruvian standard deviations greater than the Chilean ones, which in turn were substantially greater than those for the Japanese. The Hunt measures, however, did not follow this pattern. The standard deviation for High K's (H1) was greatest for Peru, as expected, but least for Chile. For the Words per T-Unit (H2) measure, the Japanese standard deviation was smallest, as expected, but the Chilean standard deviation slightly exceeded that for Peru.

In considering the above findings, it should be noted that the High K's and Words per T-Unit scores are such that a candidate at a given level of writing ability could obtain a relatively high score on either of the measures by a strategy that would lower his score on the other. Thus,

Table 6

Means, Standard Deviations, and Reliabilities of
Open-Ended Objective Measures and of Subjective Measures

| Measure | Mean | | | Standard deviation | | | Reliability | | |
|---|---|---|---|---|---|---|---|---|---|
| | Peru | Chile | Japan | Peru | Chile | Japan | Peru | Chile | Japan |
| **Open-ended objective measures** | | | | | | | | | |
| Hunt Aluminum Passage | | | | | | | | | |
| H1 High K's | 24.9 | 27.2 | 26.3 | 6.33 | 4.15 | 4.59 | .96 | .84 | .92 |
| H2 Words per T-Unit | 9.7 | 11.0 | 10.9 | 2.86 | 2.95 | 2.21 | .98 | .97 | .86 |
| (N) | (95) | (143) | (192) | | | | | | |
| Cloze Passages | | | | | | | | | |
| C1 Clozentropy score | 39.6 | 60.6 | 51.1 | 25.0 | 19.6 | 16.7 | .91 | .88 | .78 |
| C2 Standard Cloze score | 13.6 | 19.8 | 16.4 | 8.06 | 7.04 | 5.87 | .82 | .85 | .79 |
| (N) | (95) | (143) | (192) | | | | | | |
| **Subjective measures** | | | | | | | | | |
| Interview Judgments | | | | | | | | | |
| I1 Grammar | 3.09 | 3.81 | 3.48 | 1.13 | .91 | .65 | .94 | .89 | .73 |
| I2 Vocabulary | 3.12 | 3.97 | 3.56 | 1.24 | .94 | .72 | .95 | .89 | .75 |
| I3 Overall Communication | 3.18 | 4.06 | 3.61 | 1.26 | 1.00 | .76 | .94 | .89 | .74 |
| (N) | (96) | (140) | (187) | | | | | | |
| Essay Judgments | | | | | | | | | |
| E1 Content | 2.87 | 3.73 | 3.25 | 1.21 | .94 | .66 | .98 | .96 | .91 |
| E2 Form | 2.71 | 3.62 | 3.24 | 1.27 | 1.10 | .73 | .98 | .92 | .91 |
| (N) | (95) | (143) | (192) | | | | | | |

if a cautious student limited his rewriting of the Hunt passage to the introduction of a few coordinating conjunctions, he was likely to receive a High K's score at or near the maximum of 32, but a very low Words per T-Unit score of perhaps 5. The same person could instead be more daring and construct longer T-units, but at the risk of introducing ambiguities and confusions resulting in a reduction in his High K's score. This added source of variability may have had a role in the different rank-orderings in standard deviation noted earlier.

Differences between Clozentropy (C1) and Standard Cloze (C2) scores are attributable entirely to the method of scoring. Standard deviations on both measures followed the expected pattern, with those for Peru substantially larger than those for Japan. Means also followed a now-familiar pattern, greatest for Chile and lowest for Peru.

Differences in reliability, by country and by scoring method, are of particular interest. For the Clozentropy score, the reliability was .91 for Peru, .88 for Chile, and .78 for Japan, a pattern which was about as expected, given the differences in standard deviation. However, this pattern did not hold for the Standard Cloze reliabilities. Assuming that the Chilean and Japanese reliabilities for C2 were about what would be expected, the Peruvian reliability was substantially less than one would expect, particularly since the typical pattern of standard deviations held for C2 scores. The magnitude of this disparity may best be illustrated by comparing the efficiency of the two scoring methods with respect to the Peruvian subjects. Using known relationships between test length and reliability, the 50-item test would have to be expanded to more than 100 items in order for the Standard Cloze reliability of .82 to be increased

to the .91 value found for the Clozentropy score. The most likely reason for this difference is that, under the Standard Cloze scoring procedure, many items were failed by all or nearly all low-scoring subjects and, as a result, those items provided little or no discrimination among those subjects. This effect diminishes, of course, for higher-scoring groups. On the other hand, because the Clozentropy scoring procedure allows varying degrees of partial credit for many of the responses that are scored zero in the Standard Cloze procedures, it is not subject to this pronounced effect on score reliability of the subjects' ability level.

Subjective Measures

Means, standard deviations, and reliabilities of the interview and essay judgments are also given in Table 6. It is evident that for each subject group the statistics are highly similar over the interview scales shown; Grammar (I1), Vocabulary (I2), and Overall Communication (I3). Given the care taken in the judgment procedures to avoid a halo effect, it is likely that these English-speaking skills were in fact very closely related for all three groups of subjects.

As was noted for the objective measures, there was a pronounced difference in score variability, with standard deviations greatest for Peru and least for Japan. Also as noted before, the differences in reliability are attributable almost entirely to these differences in variance. Mean interview scores for Peru, Chile, and Japan were about 3, 4, and 3 1/2, respectively. The differences were sizeable in both a normative and a criterion-referenced sense, with mean differences between Chile and Peru of about one standard deviation, representing one full

level on the six-point criterion scale given in Appendix B. Reliabilities for the three interview scales averaged .94 for Peruvians, .89 for Chileans, and .74 for Japanese candidates.

It will be recalled that each subject wrote four essays, and that each of his essays was judged by eight readers, two for each of four scales, Content (Quantity), Content (Quality), Form (Quantity), and Form (Quality). By combining ratings over the four essays, there were eight ratings per candidate on each of the four scales. Because preliminary analyses of the readings showed virtually no distinction between the two Content scores, nor between the two Form scores, the scales were combined to an overall Essay Content (E1) and an overall Essay Form (E2) scale. Since the eight readers providing the Content (Quantity) ratings of each candidate's essay were independent of those giving the Content (Quality) ratings, the correlation between the two pooled sets of ratings provided a conservative procedure for estimating reliability. The reliability estimates thus obtained for overall Essay Content ratings were .98 for Peruvians, .96 for Chileans, and .91 for the Japanese. Following the same procedure, reliabilities for overall Essay Form ratings were estimated at .98, .92, and .91 for the same groups of candidates.

The pattern of means, standard deviations, and reliabilities for the essay judgments resemble those noted above for the interview judgments. Mean differences were also of the same general magnitude, with the three groups separated by intervals of about one-half standard deviation.

## Intercorrelations Among Test Scores

The relationships among scores, and in particular the intercorrelations between those which may be compared as predictors and their criteria, are central to the main purpose of this study; that is, to provide an empirical data base to facilitate the review and possible revision of the content specifications of a test of English as a foreign language.

For convenience, data will again be presented in the groupings of multiple-choice, open-ended objective, and subjective measures. To a degree this division among groups of measures, based on how they are scored, may be equated to a functional division between predictive and criterion usage. This correspondence is in part because factors of cost and feasibility favor objective measures for ultimate selection as predictors in any test that is to be administered on a large scale, and in part because subjective measures may serve better than objective measures to approximate such pragmatically important criteria as teachers' judgments about their student's ability to write well in English. However, these considerations do not necessarily rule out the use of certain objective measures, such as multiple-choice tests of reading, as criteria.

In examining the various intercorrelations, the objective TOEFL measures for the receptive skills of listening and reading will be treated as criteria, as will subjective measures of the productive language skills of speaking and writing.

The emphasis in the following discussion will be upon examining and comparing individual intercorrelation values. Given that emphasis, it would be natural to put a premium on high correlations and, of course, high correlations between predictive and criterion measures are very

desirable. However, it should be noted that ultimately we are interested in selecting a combination of measures for a test, and for that purpose, lower intercorrelations among multiple predictors are desirable. That is, each predictor should contribute score variance related to its criterion that is unique to the score variance reflected in the other predictors. In fact, a given predictive measure which correlates only modestly with the criteria may, if it taps a unique source of criterion-related variance, be given a high statistical priority for inclusion in a combined predictive test.

As may be seen in Tables 5 and 6, reliability estimates varied among the several measures and among the subject groups. The reliability of each measure influenced its correlation with other measures, and was itself influenced by test length and by the range of subjects' scores on that measure. To allow comparisons among intercorrelations not influenced by these differences, each set of observed intercorrelations is followed by a corresponding set corrected for attenuation due to unreliability.

Correlations Among Multiple-Choice Scores and Subscores

Observed intercorrelations among the nine multiple-choice scores are given in Table 7. Comparing these data by subject group, it is evident that the Peruvian correlations tended to be greatest and the Japanese ones to be smallest. "World" intercorrelations tended to fall between the Spanish and Japanese values. No general pattern over subject groups seems evident when data for one measure are compared with those for another measure.

Table 7

Observed Intercorrelations among Multiple-Choice Measures

| Measure | Subject group | Regular TOEFL T1 | T2 | T3 | T4 | T5 | Experimental TOEFL X1 | X2 | X3 | X4 |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | Peru | -- | 87 | 83 | 79 | 84 | 72 | 78 | 74 | 81 |
| Listening | Chile | -- | 83 | 65 | 62 | 73 | 68 | 61 | 65 | 61 |
| Comprehension | Japan | -- | 65 | 51 | 56 | 52 | 62 | 56 | 51 | 53 |
| | World | -- | 63 | 54 | 64 | 58 | | | | |
| T2 | Peru | 87 | -- | 76 | 75 | 88 | 70 | 77 | 71 | 79 |
| English | Chile | 83 | -- | 72 | 67 | 84 | 74 | 71 | 74 | 71 |
| Structure | Japan | 65 | -- | 62 | 62 | 63 | 60 | 65 | 61 | 57 |
| | World | 63 | -- | 68 | 67 | 77 | | | | |
| T3 | Peru | 83 | 76 | -- | 83 | 79 | 76 | 83 | 80 | 87 |
| Vocabulary | Chile | 65 | 72 | -- | 72 | 72 | 59 | 69 | 69 | 66 |
| | Japan | 51 | 62 | -- | 66 | 66 | 46 | 73 | 59 | 71 |
| | World | 54 | 68 | -- | 68 | 67 | | | | |
| T4 | Peru | 79 | 75 | 83 | -- | 79 | 81 | 84 | 83 | 83 |
| Reading | Chile | 62 | 67 | 72 | -- | 73 | 68 | 75 | 76 | 75 |
| Comprehension | Japan | 56 | 62 | 66 | -- | 61 | 55 | 73 | 70 | 73 |
| | World | 64 | 67 | 68 | -- | 67 | | | | |
| T5 | Peru | 84 | 88 | 79 | 79 | -- | 73 | 80 | 75 | 81 |
| Writing | Chile | 73 | 84 | 72 | 73 | -- | 64 | 66 | 74 | 74 |
| Ability | Japan | 52 | 63 | 66 | 61 | -- | 51 | 67 | 59 | 60 |
| | World | 58 | 77 | 67 | 67 | -- | | | | |
| X1 | Peru | 72 | 70 | 76 | 81 | 73 | -- | 83 | 85 | 77 |
| Sentence | Chile | 68 | 74 | 59 | 68 | 64 | -- | 71 | 74 | 65 |
| Comprehension | Japan | 62 | 60 | 46 | 55 | 51 | -- | 56 | 53 | 56 |
| X2 | Peru | 78 | 77 | 83 | 84 | 80 | 83 | -- | 86 | 85 |
| Words in | Chile | 61 | 71 | 69 | 75 | 66 | 71 | -- | 71 | 72 |
| Context | Japan | 56 | 65 | 73 | 73 | 67 | 56 | -- | 64 | 77 |
| X3 | Peru | 74 | 71 | 80 | 83 | 75 | 85 | 86 | -- | 85 |
| Combining | Chile | 65 | 74 | 69 | 76 | 74 | 74 | 71 | -- | 73 |
| Sentences | Japan | 51 | 61 | 59 | 70 | 59 | 53 | 64 | -- | 71 |
| X4 | Peru | 81 | 79 | 87 | 83 | 81 | 77 | 85 | 85 | -- |
| Paragraph | Chile | 61 | 71 | 66 | 75 | 74 | 65 | 72 | 73 | -- |
| Completion | Japan | 53 | 57 | 71 | 73 | 60 | 56 | 77 | 71 | -- |

Intercorrelations among the TOEFL and Experimental TOEFL scores corresponding to those shown in Table 7, but corrected for attenuation, are given in Table 8. Following the corrections for attenuation, the percentage of instances in which the Peruvian correlations exceeded those of the Chileans by more than .03 dropped from about 90 to about 50, and the percentage of instances favoring Chile over Japan dropped from about 75 to about 50. Other patterns of relationship then emerged that were not apparent from an inspection of uncorrected correlations. As might be expected, the Listening Comprehension correlations tended to be lowest, suggesting that the listening measure tapped language skills not represented in the measures presented in the written mode. This relative independence between measures in the listening and written modes was especially pronounced for the Japanese. The Japanese corrected correlations between Listening Comprehension and the written-format objective measures ranged from .61 to .81, whereas the range of those for Peru was from .78 to .94, and those for Chile, from .68 to .91. An exception to this pattern for each Hispanic group was the relatively high correlation between Listening Comprehension and English Structure (r= .94 for Peruvians and .91 for Chileans). The other exception, for Peruvians only, was the Listening Comprehension--Writing Ability correlation of .92.

The differences among corrected correlations associated with the three subject groups may well indicate real differences in the extent that the language abilities in question are less related for some foreign students than for others. A reasonable conjecture is that the relationships among component English language skills may tend to be lower for Japanese students than for those having an Indo-European language background.

Table 8

Intercorrelations among Multiple-Choice Measures, Corrected for Attenuation

| Measure | Subject group | Regular TOEFL T1 | T2 | T3 | T4 | T5 | Experimental TOEFL X1 | X2 | X3 | X4 |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 Listening Comprehension | Peru | -- | 94 | 91 | 88 | 92 | 78 | 85 | 81 | 88 |
| | Chile | -- | 91 | 73 | 73 | 82 | 80 | 68 | 77 | 69 |
| | Japan | -- | 80 | 61 | 74 | 65 | 81 | 69 | 68 | 63 |
| | World | -- | 74 | 63 | 77 | 67 | | | | |
| T2 English Structure | Peru | 94 | -- | 84 | 84 | 97 | 77 | 85 | 79 | 87 |
| | Chile | 91 | -- | 85 | 80 | 96 | 88 | 83 | 90 | 83 |
| | Japan | 80 | -- | 78 | 85 | 82 | 80 | 84 | 85 | 71 |
| | World | 74 | -- | 82 | 83 | 93 | | | | |
| T3 Vocabulary | Peru | 91 | 84 | -- | 95 | 89 | 85 | 93 | 91 | 98 |
| | Chile | 73 | 85 | -- | 90 | 86 | 74 | 85 | 88 | 80 |
| | Japan | 61 | 78 | -- | 88 | 84 | 60 | 93 | 80 | 86 |
| | World | 63 | 82 | -- | 84 | 81 | | | | |
| T4 Reading Comprehension | Peru | 88 | 84 | 95 | -- | 90 | 91 | 96 | 95 | 94 |
| | Chile | 73 | 80 | 90 | -- | 89 | 86 | 94 | 98 | 93 |
| | Japan | 74 | 85 | 88 | -- | 84 | 78 | 99 | 99 | 95 |
| | World | 77 | 83 | 84 | -- | 83 | | | | |
| T5 Writing Ability | Peru | 92 | 97 | 89 | 90 | -- | 81 | 90 | 85 | 90 |
| | Chile | 82 | 96 | 86 | 89 | -- | 78 | 79 | 91 | 88 |
| | Japan | 65 | 82 | 84 | 84 | -- | 68 | 88 | 83 | 76 |
| | World | 67 | 93 | 81 | 83 | -- | | | | |
| X1 Sentence Comprehension | Peru | 78 | 77 | 85 | 91 | 81 | -- | 92 | 95 | 85 |
| | Chile | 80 | 88 | 74 | 86 | 78 | -- | 89 | 95 | 80 |
| | Japan | 81 | 80 | 60 | 78 | 68 | -- | 77 | 77 | 73 |
| X2 Words in Context | Peru | 85 | 85 | 93 | 96 | 90 | 92 | -- | 97 | 94 |
| | Chile | 68 | 83 | 85 | 94 | 79 | 89 | -- | 90 | 88 |
| | Japan | 69 | 84 | 93 | 99 | 88 | 77 | -- | 91 | 97 |
| X3 Combining Sentences | Peru | 81 | 79 | 91 | 95 | 85 | 95 | 97 | -- | 95 |
| | Chile | 77 | 90 | 88 | 98 | 91 | 95 | 90 | -- | 92 |
| | Japan | 68 | 85 | 80 | 99 | 83 | 77 | 91 | -- | 96 |
| X4 Paragraph Completion | Peru | 88 | 87 | 98 | 94 | 90 | 85 | 94 | 95 | -- |
| | Chile | 69 | 83 | 80 | 93 | 88 | 80 | 88 | 92 | -- |
| | Japan | 63 | 71 | 86 | 95 | 76 | 73 | 97 | 96 | -- |

However, no compelling suggestion comes to mind that would explain the fact that the Peruvian correlations still tended to be somewhat greater than those for the Chilean subjects.

The Reading Comprehension measure may be considered a criterion or quasi-criterion, with some of the other objective scores treated either as estimators or as potential alternative measures. Sentence Comprehension (X1) would be expected to correlate well with Reading Comprehension (T4). As an efficient measure with face validity for testing reading comprehension, X1 could be an effective supplement to the less efficient T4 measure. Corrected correlations between X1 and T4 were high for both Spanish groups (.91 and .86 for Peruvians and Chileans), and moderate (.78) for Japanese subjects, but these were not exceptional among the adjusted correlation values. They would probably have been higher, particularly for the Japanese, if the Sentence Comprehension test had included a greater number of difficult items. There was virtually no discrimination in Sentence Comprehension scores among the more able students, because most of them missed only two or less of the thirty items.

Of particular interest is the very high relationship between Reading Comprehension (T4) and the three remaining Experimental TOEFL measures--Words in Context (X2), Combining Sentence (X3), and Paragraph Completion (X4), with corrected correlations in the .93 to .99 range. The Words in Context measure is basically a vocabulary test which provides contextual information that must be used to answer the question fully. Like Sentence Comprehension (X1), it does not go beyond the sentence level. Apparently, whatever ability the Reading Comprehension score indicates regarding beyond-sentence comprehension, that ability was closely related

to the within-sentence comprehension skills for all three groups of subjects used in this study. The Paragraph Completion (X4) measure is a multiple-choice equivalent of the Cloze task. The high relationship between it and Reading Comprehension scores gives empirical support to the suggestion that the Cloze task draws on the same underlying abilities as the more conventional tests of reading. The high correlations between Reading Comprehension and the Combining Sentence (X3) task is harder to account for. The latter calls for the ability to recognize appropriate combinations of short statements that have been embedded into a single sentence. Although decoding a structurally complex sentence is no doubt vital to effective reading, the Combining Sentences item type is not intended to tap vocabulary. Finally, Reading Comprehension and Vocabulary were highly correlated, with correlations of .95, .90, .88, and .84 for the four subject groups.

In summary, the ability to comprehend reading passages was very closely approximated, for all three subject groups, by measures of sentence comprehension, of selecting missing words for passages, of vocabulary (especially when context plays a role), and of embedding component sentences into a longer sentence semantically equivalent to its components. It may be useful to note that in all of the measures described in the above paragraph, the emphasis is on the comprehension of meaning, with little emphasis given to knowledge of standard usage, per se.

Among the nine objective measures being compared, the two which most emphasize standard usage are English Structure and Writing Ability. This may account for the very high adjusted correlations between these measures of .97 for Peru, .96 for Chile, and .93 for the World sample. For the

Japanese, the correlation was .82.

Observed and corrected intercorrelations between the subscores for the measures of Listening Comprehension, Vocabulary, and Writing Ability are given in Table 9. Corrected correlations ranging from .97 to .99 indicate that the first two Listening Comprehension scores, Sentences and Dialogues, gave nearly identical information. The Lecture subtest provided some unique score variance, as indicated by the corrected correlations between the Lecture subtest and the Sentences and Dialogues subtests, which ranged from .90 to .94 for the Hispanic students, and from .85 to .90 for the Japanese.

The two Vocabulary subtests, Sentence Completion and Synonyms, differ in that the former provides context but the latter does not. Corrected correlations between these subtests were very high for Peru and Japan at .99 and .93, respectively. For Chile, the correlation was .77.

The corrected correlations between the Writing Ability subtests, Error Recognition and Sentence Completion, were .84, .70, and .62 for Peru, Chile, and Japan. It is interesting to note that, almost without exception, these correlations between the two subtests <u>within</u> the Writing Ability section were less than those observed between Writing Ability and the other separately scored sections of TOEFL.

<u>Correlations Among Open-Ended Objective Measures and Subjective Measures</u>

As noted earlier, the interview and essay measures were developed as criteria for the productive language skills of speaking and writing. Kellogg Hunt's measure of syntactic maturity, Words per T-Unit, has been found effective for comparing English writing skills among native English

Table 9

Observed and Corrected Intercorrelations among Regular TOEFL Subsections

| TOEFL Subsection | Observed intercorrelations | | | Corrected intercorrelations | | |
|---|---|---|---|---|---|---|
| | Peru | Chile | Japan | Peru | Chile | Japan |
| **Listening Comprehension** | | | | | | |
| T1a Sentences vs. T1b Dialogues | 87 | 81 | 66 | 99 | 97 | 97 |
| T1a Sentences vs. T1c Lecture | 74 | 71 | 58 | 90 | 90 | 85 |
| T1b Dialogues vs. T1c Lecture | 77 | 70 | 57 | 94 | 91 | 90 |
| **Vocabulary** | | | | | | |
| T3a Sentence Completion vs. T3b Synonyms | 77 | 54 | 63 | 99 | 77 | 93 |
| **Writing Ability** | | | | | | |
| T5a Error Recognition vs. T5b Sentence Completion | 68 | 56 | 42 | 84 | 70 | 62 |

speakers (Hunt, 1970b). In the present discussion, then, its value as an estimator of the ability to write English as a foreign language will be examined. Since there has also been considerable interest in the utility of the Cloze procedure for assessing language skills, the Cloze task and two methods of scoring the subjects' responses will be examined in this respect.

The pattern in which the largest observed correlations generally appeared for Peru and the smallest for Japan was even more pronounced for the data in Table 10 than it was for the multiple-choice measures. Another pattern in Table 10 is that the Clozentropy correlations were generally higher than the corresponding Standard Cloze values. Also of interest are the correlations among the three interview measures and those between the Cloze scores. The remarkable similarity of summary statistics among the interview measures was noted earlier. The suggestion given then, that the interview scores--Grammar (I1), Vocabulary (I2), and Communication (I3)--are closely related, is borne out by the very high observed intercorrelations among these interview ratings. Similarly, the high observed correlations between the Clozentropy and Standard Cloze scores show that for each subject group the two scoring methods were clearly equivalent in how they rank-ordered the students. It should be recalled however that the two scoring procedures differed substantially in efficiency, particularly for low-scoring students.

The corrected correlations among the open-ended objective scores and subjective scores may be examined in Table 11. Again, the systematic differences associated with subject group did not simply drop out as a result of the correction for unreliability. In nearly all instances,

Table 10

Observed Intercorrelations among Open-Ended Objective Measures and Subjective Measures

| Measure | Subject group | Open-ended objective measures | | | | Subjective measures | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | H1 | H2 | C1 | C2 | I1 | I2 | I3 | E1 | E2 |
| H1 | Peru | -- | 30 | 51 | 49 | 57 | 56 | 56 | 52 | 52 |
| Hunt's | Chile | -- | 24 | 51 | 44 | 33 | 34 | 32 | 32 | 39 |
| High K's | Japan | -- | -04 | 34 | 30 | 24 | 26 | 27 | 21 | 27 |
| H2 | Peru | 30 | -- | 41 | 30 | 51 | 52 | 50 | 55 | 51 |
| Hunt's | Chile | 24 | -- | 49 | 52 | 42 | 46 | 43 | 42 | 43 |
| Words/T | Japan | -04 | -- | 24 | 23 | 08 | 12 | 07 | 07 | 17 |
| C1 | Peru | 51 | 41 | -- | 98 | 77 | 76 | 78 | 81 | 88 |
| Clozentropy | Chile | 51 | 49 | -- | 97 | 67 | 69 | 69 | 71 | 80 |
| | Japan | 34 | 24 | -- | 96 | 46 | 46 | 43 | 52 | 68 |
| C2 | Peru | 49 | 30 | 98 | -- | 71 | 69 | 72 | 76 | 85 |
| Standard | Chile | 44 | 52 | 97 | -- | 66 | 68 | 67 | 69 | 77 |
| Cloze | Japan | 30 | 23 | 96 | -- | 44 | 43 | 40 | 46 | 66 |
| I1 | Peru | 57 | 51 | 77 | 71 | -- | 97 | 97 | 87 | 87 |
| Interview: | Chile | 33 | 42 | 67 | 66 | -- | 95 | 95 | 75 | 83 |
| Grammar | Japan | 24 | 08 | 46 | 44 | -- | 93 | 91 | 49 | 58 |
| I2 | Peru | 56 | 52 | 76 | 69 | 97 | -- | 98 | 87 | 86 |
| Interview: | Chile | 34 | 46 | 69 | 68 | 95 | -- | 97 | 76 | 84 |
| Vocabulary | Japan | 26 | 12 | 46 | 43 | 93 | -- | 96 | 46 | 55 |
| I3 | Peru | 56 | 50 | 78 | 72 | 97 | 98 | -- | 87 | 86 |
| Interview: | Chile | 32 | 43 | 69 | 67 | 95 | 97 | -- | 77 | 83 |
| Communication | Japan | 27 | 07 | 43 | 40 | 91 | 96 | -- | 43 | 52 |
| E1 | Peru | 52 | 55 | 81 | 76 | 87 | 87 | 87 | -- | 94 |
| Essay: | Chile | 32 | 42 | 71 | 69 | 75 | 76 | 77 | -- | 89 |
| Content | Japan | 21 | 07 | 52 | 46 | 49 | 46 | 43 | -- | 76 |
| E2 | Peru | 52 | 51 | 88 | 85 | 87 | 86 | 86 | 94 | -- |
| Essay: | Chile | 39 | 43 | 80 | 77 | 83 | 84 | 83 | 89 | -- |
| Form | Japan | 27 | 17 | 68 | 66 | 58 | 55 | 52 | 76 | -- |

the Japanese correlations in Table 11 remained the smallest; about three-fourths of the Chilean correlations were smaller than the Peruvian ones. The discussion of possible reasons for these subject-group differences in correlations, given in regard to the corrected correlations of the objective scores in Table 8, applies to the data in Table 11 as well.

The pattern in which Clozentropy correlations consistently exceeded those for Standard Cloze scores disappeared once the corrections for unreliability were made. As suggested earlier, the real difference in the effect of the scoring procedures was that the Clozentropy procedure yielded greater reliability and, therefore, greater efficiency, with respect to item-writing, testing time, etc. It should be remembered, however, that the Clozentropy scoring procedure is much less efficient than the Standard Cloze procedure with respect to the time and cost of carrying out the scoring itself.

Even with corrections for attenuation, correlations between Hunt's Words per T-Unit (H2) and either of the Essay measures, E1 and E2, were only middling for the Spanish groups (.56 and .52 for Peruvians; .44 and .45 for Chileans) and very low for the Japanese (.08 and .19). The tendency for some students at a given level to favor clarity of expression (and thus attain a near-perfect "High K's" score) at the expense of using only very short T-units, and for others to do the opposite, probably explains the poor showing of Words per T-Unit as an indicator of English writing ability. Whatever the reason, correlations involving either High K's or Words per T-Unit generally showed a marked drop from Peruvian to Chilean subjects, and again from Chilean to Japanese.

Table 11

Intercorrelations among Open-Ended Objective Measures and Subjective Measures, Corrected for Attenuation

| Measure | Subject group | Open-ended objective measures H1 | H2 | C1 | C2 | Subjective measures I1 | I2 | I3 | E1 | E2 |
|---|---|---|---|---|---|---|---|---|---|---|
| H1 | Peru | -- | 31 | 55 | 55 | 60 | 59 | 59 | 54 | 54 |
| Hunt's | Chile | -- | 26 | 59 | 52 | 38 | 39 | 37 | 36 | 44 |
| High K's | Japan | -- | -04 | 40 | 35 | 29 | 31 | 33 | 23 | 30 |
| H2 | Peru | 31 | -- | 43 | 33 | 53 | 54 | 52 | 56 | 52 |
| Hunt's | Chile | 26 | -- | 53 | 57 | 45 | 49 | 46 | 44 | 45 |
| Words/T | Japan | -04 | -- | 29 | 28 | 10 | 15 | 09 | 08 | 19 |
| C1 | Peru | 55 | 43 | -- | 99 | 83 | 82 | 84 | 85 | 93 |
| Clozentropy | Chile | 59 | 53 | -- | 99 | 76 | 78 | 78 | 77 | 89 |
| | Japan | 40 | 29 | -- | 99 | 61 | 60 | 57 | 62 | 80 |
| C2 | Peru | 55 | 33 | 99 | -- | 81 | 77 | 82 | 84 | 94 |
| Standard | Chile | 52 | 57 | 99 | -- | 76 | 78 | 77 | 76 | 87 |
| Cloze | Japan | 35 | 28 | 99 | -- | 58 | 56 | 52 | 54 | 78 |
| I1 | Peru | 60 | 53 | 83 | 81 | -- | 99 | 99 | 91 | 91 |
| Interview: | Chile | 38 | 45 | 76 | 76 | -- | 99 | 99 | 81 | 90 |
| Grammar | Japan | 29 | 10 | 61 | 58 | -- | 99 | 99 | 60 | 71 |
| I2 | Peru | 59 | 54 | 82 | 77 | 99 | -- | 99 | 90 | 89 |
| Interview: | Chile | 39 | 49 | 78 | 78 | 99 | -- | 99 | 82 | 92 |
| Vocabulary | Japan | 31 | 15 | 60 | 56 | 99 | -- | 99 | 56 | 67 |
| I3 | Peru | 59 | 52 | 84 | 82 | 99 | 99 | -- | 91 | 90 |
| Interview: | Chile | 37 | 46 | 78 | 77 | 99 | 99 | -- | 83 | 92 |
| Communication | Japan | 33 | 09 | 57 | 52 | 99 | 99 | -- | 52 | 63 |
| E1 | Peru | 54 | 56 | 85 | 84 | 91 | 90 | 91 | -- | 96 |
| Essay: | Chile | 36 | 44 | 77 | 76 | 80 | 82 | 83 | -- | 95 |
| Content | Japan | 23 | 08 | 62 | 54 | 60 | 56 | 52 | -- | 84 |
| E2 | Peru | 54 | 52 | 93 | 94 | 91 | 89 | 90 | 96 | -- |
| Essay: | Chile | 44 | 45 | 89 | 87 | 90 | 92 | 92 | 95 | -- |
| Form | Japan | 30 | 19 | 80 | 78 | 71 | 67 | 63 | 84 | -- |

The two Cloze measures performed very well as indicators of essay writing ability. Appropriately, the correlations were highest with respect to Essay Form (E2), with that measure and Clozentropy correlating .93, .89, and .80 for Peruvians, Chileans, and Japanese, respectively.

Correlations Between Multiple-Choice Scores and Other Scores

The observed correlations between multiple-choice scores on the one hand, and open-ended objective scores and subjective scores on the other, will be considered next. The interview and essay measures, and to a limited degree the Cloze measures, will be regarded as criteria for selected multiple-choice measures. The observed correlations in Table 12 followed patterns generally consistent with those observed earlier. Peruvian correlations tended to be highest and Japanese correlations lowest, and Clozentropy scores usually correlated slightly more with the objective measures than did the Standard Cloze scores.

Turning to the corrected correlations in Table 13, note that for any given objective measure the correlations with the three interview scales were nearly identical. For convenience, then, Communication (I3) will be used in subsequent discussions to represent the speaking criteria.

The Communication measure came nearer than any other in the study to serving as a criterion for Listening Comprehension. The correlations were gratifyingly high, with the .82 observed for the Japanese being the largest correlation for that language group between spoken communication and the several objective measures. The second best predictor of I3 for the Japanese was English Structure ($r = .71$). For Peruvian and Chilean subjects, the corrected correlations between Listening Comprehension

Table 12

Observed Correlations between Multiple-Choice Measures, and Open-Ended Objective Measures and Subjective Measures

| Multiple-choice measure | Subject group | Open-ended objective measures: Hunt psg. H1 High K's | Hunt psg. H2 Wds/ T | Cloze psgs. C1 Cloz-entr. | Cloze psgs. C2 Std. Cloze | Subjective measures: Interview I1 Gram-mar | Interview I2 Vo-cab. | Interview I3 Com-mun. | Essays E1 Con-tent | Essays E2 Form |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | Peru | 45 | 43 | 81 | 80 | 79 | 79 | 79 | 80 | 87 |
| Listening | Chile | 34 | 34 | 68 | 65 | 69 | 68 | 71 | 72 | 77 |
| Comprehension | Japan | 20 | 12 | 52 | 51 | 66 | 66 | 65 | 52 | 63 |
| T2 | Peru | 44 | 43 | 83 | 78 | 80 | 81 | 81 | 82 | 87 |
| English | Chile | 36 | 47 | 82 | 79 | 78 | 77 | 77 | 81 | 89 |
| Structure | Japan | 26 | 09 | 64 | 62 | 53 | 53 | 54 | 46 | 68 |
| T3 | Peru | 57 | 42 | 82 | 83 | 75 | 76 | 75 | 74 | 78 |
| Vocabulary | Chile | 35 | 36 | 70 | 70 | 64 | 64 | 62 | 64 | 70 |
| | Japan | 25 | 31 | 64 | 64 | 43 | 49 | 46 | 39 | 57 |
| T4 | Peru | 56 | 47 | 79 | 79 | 79 | 79 | 78 | 77 | 78 |
| Reading | Chile | 44 | 47 | 76 | 77 | 60 | 62 | 61 | 56 | 67 |
| Comprehension | Japan | 29 | 18 | 69 | 65 | 44 | 45 | 44 | 48 | 58 |
| T5 | Peru | 45 | 46 | 82 | 80 | 79 | 78 | 79 | 79 | 87 |
| Writing | Chile | 38 | 50 | 76 | 76 | 69 | 68 | 66 | 71 | 78 |
| Ability | Japan | 28 | 22 | 64 | 63 | 44 | 47 | 45 | 53 | 61 |
| X1 | Peru | 64 | 44 | 82 | 79 | 80 | 79 | 80 | 77 | 78 |
| Sentence | Chile | 46 | 36 | 77 | 73 | 64 | 67 | 66 | 70 | 74 |
| Comprehension | Japan | 32 | 03 | 55 | 51 | 47 | 47 | 48 | 38 | 50 |
| X2 | Peru | 55 | 46 | 87 | 84 | 78 | 77 | 79 | 78 | 82 |
| Words in | Chile | 36 | 45 | 79 | 78 | 69 | 72 | 71 | 67 | 76 |
| Context | Japan | 27 | 29 | 73 | 69 | 47 | 50 | 47 | 46 | 61 |
| X3 | Peru | 65 | 40 | 82 | 81 | 78 | 78 | 79 | 75 | 78 |
| Combining | Chile | 41 | 45 | 81 | 80 | 60 | 64 | 62 | 65 | 72 |
| Sentences | Japan | 32 | 16 | 65 | 64 | 42 | 43 | 39 | 47 | 64 |
| X4 | Peru | 57 | 42 | 89 | 89 | 73 | 73 | 74 | 76 | 82 |
| Paragraph | Chile | 35 | 52 | 77 | 77 | 58 | 59 | 58 | 59 | 68 |
| Completion | Japan | 35 | 23 | 72 | 70 | 43 | 48 | 46 | 45 | 58 |

Table 13

Correlations between Multiple-Choice Measures and Open-Ended Objective Measures and Subjective Measures, Corrected for Attenuation

| Multiple-choice measure | Subject group | Open-ended objective measures: Hunt psg. H1 High K's | Hunt psg. H2 Wds/T | Cloze psgs. C1 Cloz-entr. | Cloze psgs. C2 Std. Cloze | Subjective measures: Interview I1 Gram-mar | Interview I2 Vo-cab. | Interview I3 Com-mun. | Essays E1 Con-tent | Essays E2 Form |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | Peru | 47 | 45 | 88 | 91 | 84 | 84 | 84 | 83 | 91 |
| Listening | Chile | 38 | 36 | 75 | 73 | 76 | 75 | 78 | 76 | 83 |
| Comprehension | Japan | 23 | 14 | 64 | 62 | 84 | 83 | 82 | 59 | 72 |
| T2 | Peru | 47 | 45 | 91 | 90 | 86 | 87 | 87 | 86 | 92 |
| English | Chile | 42 | 50 | 93 | 91 | 88 | 87 | 87 | 88 | 98 |
| Structure | Japan | 31 | 11 | 82 | 79 | 70 | 69 | 71 | 55 | 81 |
| T3 | Peru | 62 | 45 | 92 | 98 | 82 | 83 | 82 | 80 | 84 |
| Vocabulary | Chile | 43 | 41 | 83 | 86 | 77 | 77 | 73 | 74 | 81 |
| | Japan | 29 | 38 | 80 | 80 | 55 | 62 | 59 | 45 | 66 |
| T4 | Peru | 62 | 51 | 89 | 94 | 88 | 87 | 87 | 84 | 85 |
| Reading | Chile | 56 | 56 | 91 | 97 | 74 | 76 | 73 | 67 | 79 |
| Comprehension | Japan | 36 | 23 | 94 | 88 | 62 | 62 | 62 | 61 | 73 |
| T5 | Peru | 49 | 49 | 91 | 94 | 86 | 85 | 86 | 85 | 93 |
| Writing | Chile | 45 | 55 | 87 | 89 | 79 | 78 | 75 | 77 | 88 |
| Ability | Japan | 33 | 27 | 83 | 81 | 59 | 62 | 60 | 64 | 73 |
| X1 | Peru | 68 | 47 | 90 | 91 | 87 | 85 | 87 | 81 | 83 |
| Sentence | Chile | 56 | 41 | 92 | 89 | 76 | 80 | 79 | 80 | 87 |
| Comprehension | Japan | 40 | 04 | 74 | 68 | 65 | 64 | 66 | 47 | 62 |
| X2 | Peru | 59 | 49 | 96 | 98 | 85 | 83 | 86 | 83 | 87 |
| Words in | Chile | 43 | 52 | 94 | 96 | 83 | 86 | 84 | 78 | 88 |
| Context | Japan | 32 | 36 | 95 | 89 | 63 | 66 | 63 | 55 | 73 |
| X3 | Peru | 71 | 43 | 92 | 95 | 86 | 85 | 87 | 81 | 84 |
| Combining | Chile | 51 | 52 | 99 | 99 | 73 | 78 | 75 | 76 | 86 |
| Sentences | Japan | 41 | 21 | 90 | 88 | 60 | 61 | 56 | 61 | 83 |
| X4 | Peru | 61 | 45 | 98 | 99 | 79 | 79 | 80 | 81 | 87 |
| Paragraph | Chile | 43 | 60 | 90 | 95 | 70 | 71 | 68 | 69 | 78 |
| Completion | Japan | 40 | 27 | 90 | 86 | 55 | 61 | 59 | 52 | 67 |

and spoken Communication (I3) were .84 and .78, respectively. For both Hispanic groups, a number of the objective measures presented in the written mode correlated more highly with I3 than did the Listening Comprehension measure. For Peruvians, there was a near five-way tie among the objective measures for estimating spoken Communication. Measures T2, T4, X1, and X3 correlated with the I3 criterion at .87, and Words in Context (X2) did so at .86. For Chileans, the best estimator of spoken Communication was English Structure, correlating at .87, closely followed by Words in Context at .84.

For estimating Essay Form (E2) scores, the English Structure test again performed particularly well, yielding corrected correlations of .92, .98, and .81 for Peru, Chile, and Japan, respectively. Writing Ability and Words in Context also performed very satisfactorily in estimating E2 scores.

There is an increasing body of literature suggesting that the Cloze procedure is effective for estimating reading comprehension. Considering the Reading Comprehension (T4) measure as a criterion, that contention is strongly supported by the present data, with corrected correlations averaged over the two scoring procedures of about .91 for Peruvians, .95 for Chileans, and .91 for Japanese.

As noted early in this report, the advantages of the Cloze procedures are to some degree offset by the problem of administering and scoring open-respons. tests on a large scale, as compared to using multiple-choice measures. I , therefore, of interest to treat Cloze scores as criteria, to see whether measures that are more readily obtained and scored can be used in their place. Correlations between the Paragraph Completion

measure, which is a multiple-choice variant of the Cloze task, and Cloze scores (C1 and C2) were gratifyingly high, at .98 and .99 for Peruvians, .90 and .95 for Chileans, and .90 and .88 for the Japanese. Interestingly, the Words in Context (X2) and Combining Sentences (X3) tasks also correlated very highly with the Cloze scores. Correlations between X2 and the Cloze scores averaged about .97, .95, and .92 for Peru, Chile, and Japan, respectively. Those between X3 and the two Cloze scores averaged about .93, .99, and .89, for subject groups in the same order.

## Evaluations of Multiple-Choice and Open-Ended Objective Measures

A central purpose of the study was to make comparisons among a wide variety of measures, both objective and subjective, to serve as either predictors or criteria that would be of use in specifying the content of an "ideal" TOEFL. To further this purpose, each of the multiple-choice and open-ended objective measures will now be considered in turn, with information regarding its relative merits for use in an operating TOEFL. This information will be taken from the preceding discussion of results.

### Multiple-choice measures

Measures under immediate consideration for their desirability as part of a future TOEFL are those now in TOEFL and the additional objective measures found in the Experimental TOEFL developed for the present study.

T1: Listening Comprehension. Criterion measures of listening comprehension were not developed as part of the study, nor were alternative experimental measures to the three subtests making up the TOEFL Listening Comprehension section. However, the study does provide a basis for some conclusions about the usefulness of this section and its component parts.

Evidence that the listening measures contributed variance not found when materials are presented in the written mode is available in Table 7, where generally lower correlations were found between Listening Comprehension and the eight written-mode objactive measures than among the latter. Furthermore, the finding of marked reductions in item difficulty associated with presenting equivalent items entirely in the written mode in the Sentence Comprehension (X1) measure, as opposed to presenting them in the spoken mode in Listening Comprehension/Sentences (T1a), suggests that much of the difficulty in the latter is indeed attributable to the listening task itself, rather than to such factors as reasoning ability, general vocabulary, and reading ability.

Further evidence that the Listening Comprehension (T1) measure is working satisfactorily are the correlations between it and the Interview-Communication (I3) scores shown in Tables 12 and 13. The Listening Comprehension measure was an effective estimator of spoken communication ability for the Hispanic groups, and was the best estimator of this ability among the objective measures for the Japanese.

Information regarding the relationships among the three component measures of Listening Comprehension was also obtained. As shown in Table 9, the Sentences and Dialogues components were so highly intercorrelated as to be virtually interchangeable, but the Lecture component contributed some unique variance. For Peruvian and Chilean subjects, the Sentences and Dialogues portions estimated spoken Communication (I3) equally well, with corrected correlations for both subtests of about .88 for Peruvians and .78 for Chileans. The Lecture portion correlated with I3 to a much lesser degree, at about .71 for both groups. For the Japanese, spoken

Communication was estimated by T1a, T1b, and T1c at corrected correlation values of .82, .75, and .77, respectively, suggesting that the Sentences (T1a) component was the most effective estimator of I3 scores.

T2: English Structure. This measure appears superior in several respects. One of the shortest objective tests (20 minutes), it ranked second among the nine in reliability. As shown in Table 13, English Structure was the best all around estimator of Essay Form, and for the two Spanish groups it vied with Words in Context as the best estimator of Interview-Communication scores. As shown in Table 8, English Structure "clustered" with Writing Ability, perhaps because of an emphasis on "standard usage" skills, with the two measures having corrected correlations of .97, .96, and .82 for Peru, Chile, and Japan respectively.

T3: Vocabulary. This measure is also efficient. A 15-minute test, its reliability is greater than that of the 40-minute Reading Comprehension section. Some of its critics have objected to this measure on the grounds that vocabulary testing may encourage a tendency they believe is already too prevalent among students learning a foreign language--that of studying words in isolation, with the implied neglect of other aspects of language learning. Actually, the Vocabulary measure has two component parts, Vocabulary-Sentence Completion (T3a) and Vocabulary-Synonyms (T3b). The former provides context but the latter does not, and is thus particularly vulnerable to the criticism noted above.

For Peru and Japan, the two Vocabulary components worked about equally well in estimating Reading Comprehension scores and Essay Form scores. Among Chileans, however, the Vocabulary-Sentence Completion (T3a) measure was the better estimator of both criteria.

Using Interview-Vocabulary measure I2 as a criterion, the two written Vocabulary components (T3a and T3b) were equally effective for the Peruvian subjects, with corrected correlations of .83 and .83. For Chileans and Japanese, however, Sentence Completion (T3a) was clearly a better estimator of Interview-Vocabulary than was the Synonyms (T3b) component. the corrected correlations between I2 and T3a were .79 and .74 for the latter two subject groups, compared to I2 versus T3b correlations of .66 and .54 for the same groups.

T4: Reading Comprehension. Although criterion measures of reading comprehension were not developed as part of the study, the study provided very useful information about measuring this ability. Note first that the Reading Comprehension section is the least efficient of those appearing in TOEFL; although it is one of the two longest sections, it has the lowest reliability. Furthermore, Reading Comprehension items are expensive to develop and difficult to revise for inclusion in a final test, following pretesting. Bearing these problems in mind, it is interesting to note in Table 8 the very high corrected correlations between Reading Comprehension (T4) and Experimental TOEFL measures X2, X3, and X4 of .94 to .96 for Peru, .93 to .98 for Chile, and .95 to .99 for Japan. If the Sentence Comprehension (X1) section had included items at a higher level of difficulty, the same would probably have been observed for that item type as well. The implications regarding the use of one or more of these item types to test reading comprehension will be considered as each is discussed in turn. As may be seen in Table 13, Reading Comprehension scores are also closely approximated by the two Cloze scores.

T5: Writing Ability. The Essay Form measure was specifically developed as a writing ability criterion. As indicated in Table 13, the Writing Ability section estimated that criterion very well, but only to about the same degree as did the Words in Context measure. For the Chileans and Japanese, the Writing Ability section was substantially outperformed by English Structure as an estimator of actual writing performance.

The component subtests of Writing Ability--Error Recognition and Sentence Completion--were less highly interrelated than other sets of tests, with corrected correlations of .84 .70, and .62 for Peru, Chile, and Japan. These correlations are lower than most of the values observed between measures yielding separately reported scores, such as Writing Ability and Listening Comprehension. Of the two Writing Ability subtests, Error Recognition was superior for estimating the Essay Form criterion for all three subject groups. Corrected correlations between Error Recognition (T5a) and Essay Form (E2) were .93, .90, and .75, while those between Sentence Completion (T5b) and Essay Form (E2) were only .84, .70, and .55 for the three subject groups.

As noted above, English Structure and Writing Ability are highly correlated, with the two measures perhaps constituting a "standard usage" cluster. This would suggest that if both measures are retained, they might be used jointly to provide a single reported score rather than separate scores.

Experimental Section X1: Sentence Comprehension. This measure is fully equivalent to the first portion of the TOEFL Listening section, Sentence Comprehension (T1a), except that it is presented in the written

rather than the spoken mode. It is an efficient, highly reliable measure. However, the Sentence Comprehension (X1) test was too easy. The large number of participants receiving perfect or near-perfect scores depressed the correlations between Sentence Comprehension and other measures to some extent, especially for the Japanese. If it is to be used in future examinations, items should be introduced that cover a broader range of difficulty.

As the only measure of reading comprehension at the very important sentence level, the Sentence Comprehension item type has much to recommend it. Even in its present easy form, the measure correlates well with Reading Comprehension scores. It has the further advantage of face validity to recommend its use as an alternative or supplement for the more cumbersome Reading Comprehension measure.

Experimental Section X2: Words in Context. This is a vocabulary measure in which words and expressions in sentence context provide the stimuli. By testing vocabulary in context, the X2 item format reduces the basis for the criticism that vocabulary testing may foster the study of words in isolation. Actually, each item requires the use of two sources of information to answer it fully--the meaning of the underlined word or expression, and the added, interacting contextual information. The X2 measure, as developed, emphasizes "communicativeness" or the recognition or generating of equivalent messages, rather than "well-formedness" or knowledge of standard usage.

Corrected correlations between the Words in Context measure and Reading Comprehension were .96, .94, and .99 for the three subject groups. The correlations between X2 and Cloze scores C1 and C2 were almost as

high, averaging .97, .95, and .92 for Peruvians, Chileans, and Japanese, respectively (see Table 13). Thus, Words in Context provides an excellent alternative to either the Reading Comprehension or the Cloze scores, both of which are considerably less efficient and more costly to use. It does not, however, have as much face validity for estimating reading ability as has the present Reading Comprehension measure. The Words in Context (X2) measure is substantially more related to the Reading Comprehension and Cloze measures than is Vocabulary, which suggests that X2 does indeed require the use of context in a meaningful way.

Experimental Section X3: Combining Sentences. The Combining Sentences measure was developed as a multiple-choice approximation to the skills presumably tapped by Kellogg Hunt's Words per T-Unit (H2) measure. None of the objective measures, including Combining Sentences, correlated well with Words per T-Unit. However, X3 was the best of the objective estimators of H1, the number of "K's" effectively expressed.

The Combining Sentences measure correlated well with Reading Comprehension and with the Cloze scores. However, it is difficult to produce these Combining Sentences items, and to respond to them, so the measure seems less promising than the others included in the Experimental TOEFL.

Experimental Section X4: Paragraph Completion. The Paragraph Completion measure was developed as a multiple-choice equivalent of the Cloze measure. Corrected correlations between X4 and Cloze scores C1 and C2 were high, but not quite as high as those between Words in Context and the Cloze scores. The Paragraph Completion measure was also highly correlated with Reading Comprehension, but again not quite at the level observed for Words in Context.

An advantage of the Paragraph Completion format for estimating either Cloze scores or Reading Comprehension is the face validity associated with the fact that it consists of passages of connected prose, rather than a set of distinct, unrelated sentences.

Open-Ended Objective Measures

The measures derived from the Hunt and Cloze procedures are in some respects quasi-criteria, for which attempts were made to develop multiple-choice equivalents, yet they are not fully established as criteria in their own right. Neither set of measures is readily amenable to testing on a large scale, although some of the scores for each can be generated objectively.

H1 and H2: Hunt Aluminum Passage Scores. The Hunt task of rewriting a passage presented in the form of 32 very short "kernel" sentences has been found effective in rank-order Americans on what is, presumably, "syntactic maturity" (Hunt, 1970b). Particularly useful for that purpose is the "Words per T-Unit" measure, a refinement of the traditional sentence-length index of language complexity. In comparison with other measures used in the present study, the performance of the Hunt measures in estimating essay scores was disappointing (see Table 11). The corrected correlations between Words per T-Unit (H2) and Essay Form (E2) were .52, .45, and .19 for Peru, Chile, and Japan, respectively, as compared, for example, with Clozentropy (C1) versus Essay Form (E2) correlations of .93, .89, and .80, and with English Structure (T2) versus Essay Form (E2) correlations of .92, .98, and .81 for the same subject groups.

It may be that the correlations involving Words per T-Unit are comparatively low for foreign students because "syntactic maturity" is confounded with other factors. For example, among foreign students at a given level of "syntactic maturity" in English, individual differences in risk-taking behavior may result in a wide range of structural complexity in their responses to the Hunt rewriting task. The considerable range and variability in the High K's scores lend support to this interpretation.

The Hunt task remains intriguing as an open-ended measure of English structure little affected by vocabulary or creativity in essay writing. An alternative scoring procedure, combining level of clarity (H1) with level of complexity (H2) was carried out, with results very similar to those observed for H2 alone.

C1 and C2: Cloze Passage Scores. In conjunction with either the Standard Cloze or the Clozentropy scoring procedure, the Cloze task performed very satisfactorily, particularly for estimating the Essay Form subjective criterion measure and the Reading Comprehension multiple-choice criterion measure (see Tables 11 and 13). Performance on the Cloze task was, in turn, very well estimated by several of the multiple-choice measures, including Words in Context, Paragraph Completion, and to a slightly lesser degree, English Structure. It will be recalled that the Paragraph Completion (X4) measure was designed with the express purpose of approximating the Cloze scores.

The Clozentropy scoring procedure was substantially more efficient than the Standard Cloze procedure for low-scoring subjects. Except for that difference, the scores were functionally very similar.

It is interesting to note that, as one would expect, the Clozentropy correlations for the written subjective criterion, Essay Form (E2), were substantially higher than those for the spoken criterion, Communication (I3). The rank-ordering of corrected correlations between Clozentropy (C1) scores and the five TOEFL subscores is also logical, for the Chilean and Japanese data (see Table 13). Among Chileans, the TOEFL subscores most highly correlated with Clozentropy scores were English Structure and Reading Comprehension (r = .91 to .93); next were Writing Ability and Vocabulary (r = .83 to .87); the lowest correlation was for Listening Comprehension (r = .75). Among the Japanese, the highest TOEFL subtest correlation with Clozentropy scores was observed for Reading Comprehension (r = .94); intermediate correlations were observed for English Structure, Vocabulary, and Writing Ability (r = .80 to .83); once again, the lowest value was for Listening Comprehension (r = .64). Corrected correlations between Clozentropy and TOEFL subscores for Peruvians were so similar, ranging only from .88 to .92, that comparisons of rank order are not meaningful. Rank-orderings of uncorrected correlations between Clozentropy scores and TOEFL subscores (see Table 12) followed essentially the same patterns. Among Chileans, English Structure again ranked highest and Listening Comprehension lowest; among Japanese, Reading Comprehension and Listening Comprehension again ranked highest and lowest, respectively; for Peruvians, the range of correlations was again too small (.79 to .83) to meaningfully compare rank-orderings.

The above data conflict with those of Darnell (1970), who found that Clozentropy scores correlated highest with the TOEFL Listening Comprehension scores, and second highest with Vocabulary. It should be noted that

the difference he observed for correlations between Clozentropy and the two TOEFL scores just noted was minute (.736 versus .733), particularly given a sample of only 48 students. The fact remains, however, that among the TOEFL subtests, Listening Comprehension ranked highest in correlation with Clozentropy scores in the Darnell study and lowest in the present study. This difference is of interest, and an explanation should be sought.

## DISCUSSION

Limitations of the conclusions will be considered first. The implications of the study will then be discussed with regard to possible applications for the TOEFL program.

### Limitations

In considering the results of this study, it is useful to keep in mind the following limitations.

First, the study was restricted to evaluating and comparing TOEFL and other measures for the assessment of the foreign student's present skills in English as a second language that are considered important for his success in an American college or university. Thus it did not include the systematic use of other inputs such as previous grade-point average or verbal and mathematical aptitude measured with tests in the student's native language.

Second, because the study was primarily correlational, the results cannot be interpreted causally. They may show, for example, certain differences in patterns of test scores associated with differences in the subjects' native languages, but they cannot be used as evidence that these score differences are necessarily due to language background differences.

Third, the study was focused on comparisons among item formats, and was not directly concerned with differences in item content within a given item format, which can be substantial. Items in the English Structure format, for example, may differ considerably in the emphasis given to standard usage. A new English Structure subtest differing in emphasis on

standard usage could yield quite different results, particularly if the candidate groups differed in the emphasis given to formal English usage in their learning of English as a second language.

A final caveat has to do with the great variety of linguistic, cultural, and other background variables associated with the TOEFL candidate population. Differences appearing among the three groups participating in this study only underscore the importance of this observation. The inclusion of subject groups of very different linguistic and cultural backgrounds made the results more generalizable than they would otherwise have been, but there remain many subject groups, such as Pakistanis, Israelis, and Turks, for whom certain findings may be very different from those obtained with Peruvian, Chilean, and Japanese subjects.

## Implications for TOEFL Content Specifications

In this study nine TOEFL item formats and four alternative formats using multiple-choice questions have been evaluated for possible use in a revised TOEFL. The implications of the study for each format will first be examined. This will be followed by some suggestions regarding the content specifications that might be used for a revised TOEFL.

Section T1, Listening Comprehension, was found to be relatively independent of the other objective measures, and to correlate well with spoken Communication (I3). These findings add empirical support to the logical reasons for retaining Listening Comprehension as a separate measure. Although the Sentences (T1a) and Dialogues (T1b) components are highly interdependent, face validity considerations would suggest that both of these as well as the Lecture (T1c) component, be retained in the Listening Comprehension section.

Section T2, English Structure, correlated highly with the language production criteria, i.e., spoken Communication (I3) and Essay Form (E2), as well as with Writing Ability (T5). The implication of the first two correlations is that the English Structure format clearly should be retained. The third correlation suggests that English Structure and Writing Ability scores could well be combined.

Section T3, Vocabulary, was observed to be highly efficient, requiring only 15 minutes testing time to yield a satisfactory reliability. Within the Vocabulary section, the Sentence Completion format was superior to the Synonyms format for estimating Reading Comprehension (T4), Essay Form (E2), and spoken Vocabulary (I2). The Synonyms format, therefore, should probably be dropped, particularly since it is already suspected of increasing the tendency to study words in isolation. The very high correlations between Vocabulary (T3) and Reading Comprehension (T4) suggest that these scores, too, should be combined.

Section T5, Writing Ability, correlated with Essay Form (E2) with values of .93, .88, and .73 (corrected for attenuation) for Peru, Chile, and Japan. As noted earlier, the Error Recognition (T5a) component was considerably more valid than the Sentence Completion (T5b) part for all three subject groups, having correlations with the Essay Form (E2) criterion of .93, .90, and .75, respectively. The suggestion that the Writing Ability section (or its Error Recognition subsection) be replaced by an actual writing sample received little statistical support from the Peruvian and Chilean data, but received rather more support from the Japanese data. The suggestion, of course, has pedagogical as well as statistical bases, and the data presented here apply only to the latter.

The relationships between Writing Ability and English Structure scores, and its implication was noted on page 82.

Section X1, (written) Sentence Comprehension, proved very easy, and this reduced its correlations with other measures. As an efficient measure of reading comprehension at the sentence level, it is worthy of further study. A Sentence Comprehension section covering an adequate range of item difficulties could be developed and tried out for possible inclusion in a revised TOEFL. Increasing the difficulty of Sentence Comprehension items could be readily accomplished by using more difficult statements or questions, and by introducing answer choices that call for finer discriminations.

Section X2, Words in Context, showed extremely high correlations with Reading Comprehension and Cloze scores. An implication of the former is that the Words in Context format could be used to supplement the less efficient Reading Comprehension measure.

Section X3, Combining Sentences, was relatively difficult to develop, and it was outperformed by other measures as an estimator of various criterion scores. It should be dropped from consideration for inclusion in a revised TOEFL.

Section X4, Paragraph Completion, was also outperformed by other multiple-choice measures, even for estimating Cloze scores. It, too, should probably be dropped from present consideration for TOEFL.

From the above discussion, it would appear that four of the measures studied would not be likely prospects for inclusion in a revised TOEFL. These are the Vocabulary/Synonyms (T3b) and Writing Ability/Sentence Completion (T5b) subtests from TOEFL, and the Combining Sentences (X1) and

Paragraph Completion (X4) measures from Experimental TOEFL. The Reading Comprehension measure might be reduced from 40 to perhaps 20 or 30 minutes, and supplemented by another measure having face validity for reading comprehension, such as Sentence Comprehension (X1). The status of Vocabulary/Sentence Completion (T3a) is uncertain.

## Implications for the Number of Scores to Report

The question, "How many scores should be reported for TOEFL?" was raised at the beginning of this paper, and it was noted that the answer depended not only on the logical distinctions among component skills, but on how independent these skills are, in fact, for foreign students. Later in the paper it was observed that the Listening Comprehension subtests are relatively independent of the other multiple-choice measures, and that the English Structure and Writing Ability measures form one cluster and the Reading Comprehension and Vocabulary measures form another. These and other considerations suggest a revised TOEFL having several components, but yielding only three scores: I. Listening Comprehension, II. English Structure and Writing Ability, and III. Reading Comprehension and Vocabulary in Context.

All of the above comments regarding implications for TOEFL content specifications derive from the data gathered in the study, with some additional consideration given to face validity. As such, they are subject to the limitations of the study noted earlier. Furthermore, any consideration of these or other suggestions regarding possible changes in TOEFL specifications must also take into account such questions as cost, timing, and the acceptance of the changes by TOEFL score users.

## REFERENCES

Bilyeu, E. E. The Pimsleur Modern Language Proficiency Tests: Spanish Proficiency Tests. Journal of Educational Measurement, 1969, 6, 48-50.

Carroll, J. B. The psychology of language testing. In Alan Davies (Ed.), Language testing symposium: A psycholinguistics approach. London: Oxford University Press, 1968. Chap. 4. Pp. 46-69.

Carroll, J. B., Carton, A. S., & Wilds, C. P. An investigation of "cloze" items in the measurement of achievement in foreign languages. Cambridge, Mass.: Laboratory for Research in Instruction, Graduate School of Education, Harvard University, April 1959. ERIC Doc. ED 021513.

Darnell, D. K. Clozentropy: A procedure for testing English language proficiency of foreign students. Speech Monographs, March 1970, 37(1), 36-46.

Fisher, W. D., & Masia, B. B. Review of Modern Language Aptitude Test. In O. K. Buros (Ed.), The Sixth Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press, 1965.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. Research Monograph No. 6. New York: College Entrance Examination Board, 1966.

Hunt, K. W. Do sentences in the second language grow like those in the first? Paper presented at the TESOL convention, San Francisco, March 1970. (a)

Hunt, K.W. Syntactic maturity in school children and adults. Society for Research in Child Development. Monograph, 1970, 35 (1, Whole No. 134), 1-67. (b)

Manual for Peace Corps Language Testers. Unpublished document. Princeton, N.J.: Educational Testing Service.

Manual for TOEFL Score Recipients. Princeton, N. J.: College Entrance Examination Board and Educational Testing Service, 1973.

Oller, J. W. Cloze tests of second language proficiency and what they measure. Language Learning, 1973, 23, 105-118.

Pitcher, B., & Ra, J. B. The relation between scores on the Test of English as a Foreign Language and the ratings of actual theme writing. Statistical Report 67-9. Princeton, N. J.: Educational Testing Service, 1967.

Reilly, R. R. A note on Clozentropy: A procedure for testing English language proficiency of foreign students. Speech Monographs, 1971, 38, 350-353.

Rice, F. A. The Foreign Service Institute tests language proficiency. Linguistic Reporter, 1959, 1(2), 4.

Swineford, F. Test analysis--Test of English as a Foreign Language. Statistical Report 71-112. Princeton, N.J.: Educational Testing Service, 1971.

Taylor, W. L. Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 414-438.

Taylor, W. L. Recent developments in the use of "Cloze Procedure." Journalism Quarterly, 1956, 33, 42-48.

Test of English as a Foreign Language: Interpretive Information. Princeton, N.J.: College Entrance Examination Board and Educational Testing Service, 1970.

Wilds, C. P. Language proficiency testing: Language testing in the Peace Corps. Undated, 8 pages.

APPENDIX A: Materials for Scoring Hunt's Aluminum Passage

I. Score Sheet

ID # _____ SUBJECT _____ SCORER _____

| Sent. | T-Units | Words |
|---|---|---|
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |
| 15 | | |
| 16 | | |
| 17 | | |
| 18 | | |
| 19 | | |
| 20 | | |
| TOTALS | | |

| K | T-Unit | Adequacy of K | | | |
|---|---|---|---|---|---|
| | | High | Med | Low | A |
| 1. Aluminum is a metal | | | | | |
| 2. It is abundant | | | | | |
| 3. It has many uses | | | | | |
| 4. It comes from bauxite | | | | | |
| 5. Bauxite is an ore | | | | | |
| 6. B. looks like clay | | | | | |
| 7. B. contains aluminum | | | | | |
| 8. It contains several other substances | | | | | |
| 9. Workmen extract these other substances from the B. | | | | | |
| 10. They grind the B. | | | | | |
| 11. They put it in tanks | | | | | |
| 12. Pressure is in the tanks | | | | | |
| 13. The other substances form a mass | | | | | |
| 14. They remove the mass | | | | | |
| 15. They use filters | | | | | |
| 16. A liquid remains | | | | | |
| 17. They put it through several other processes | | | | | |
| 18. It finally yields a chemical | | | | | |
| 19. The chemical is powdery | | | | | |
| 20. It is white | | | | | |
| 21. The chemical is alumina | | | | | |
| 22. It is a mixture | | | | | |
| 23. It contains aluminum | | | | | |
| 24. It contains oxygen | | | | | |
| 25. Workmen separate the aluminum from the oxygen | | | | | |
| 26. They use electricity | | | | | |
| 27. They finally produce a metal | | | | | |
| 28. The metal is light | | | | | |
| 29. It has a luster | | | | | |
| 30. The luster is bright | | | | | |
| 31. The luster is silvery | | | | | |
| 32. This metal comes in many forms. | | | | | |
| TOTALS | | | | | |

APPENDIX A - Continued

II. Abbreviated Instructions for Scoring Aluminum Passage Protocol

A. Mark the end of each T-unit with a red slash.

1. Hunt (p.4) defines the T-unit as "one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it."

2. For the purpose of scoring, each sentence, even if it is a fragment, is assumed to contain at least one T-unit. Thus, you will always indicate a T-unit boundary at the end of each sentence.

3. Within-sentence T-unit boundaries will occur whenever two adjoining clauses are independent. If, on the other hand, one of the adjoining clauses is subordinate to the other, they are both part of the same T-unit.

4. Writers may use an inappropriate word to join two clauses. In that case, make the best judgment based on the sense of the passage, of whether the clauses are independent. Example (#7): "It has many uses why it is abundant..." Here, the "why" seems to imply subordination, as though "because" had been used. Thus, the two phrases are considered part of the same T-unit.

5. Extraneous T-units (those not involving at least one K) are treated the same as other T-units. Note them at the bottom of the score sheet, however.

6. T-unit fragments will not ordinarily occcur, because a given T-unit includes any attached or embedded nonclausal structure (A-1). They can occur in at least two instances, however.

   a. When a segment punctuated as a sentence is actually less than a clause.

   b. When there is an incomplete sentence, which clearly includes a complete T and a T-fragment.

Indicate T-unit fragments at the bottom of the score sheet.

APPENDIX A - Continued

B. Evaluate the "Adequacy of K"

1. After each K, check one of the 4 categories: Hi, Med, Low, or Abs. The meaning of each category is roughly as follows (consult B-2 for details):

   a. Hi. Check this category if the K is clearly stated. It does not need to be stated in standard English.

   b. Med. Check this category if the information of K seems to have been correctly presented, but with some ambiguity. A useful rule of thumb is to mark the K down to "Med" if careful reading was required before the meaning was clear.

   c. Low. If a K is ambiguously stated, or is misstated, or is misleading, mark it "Low". Especially in cases of poor writing, it may be difficult to decide between "Low" and "Abs." Here, the basic rule is to judge whether the writer attempted to express the K. If the judgment is "yes," then check that K as "Low," not as "Abs."

   d. Abs. If the K was not referred to in any T or combination of T's check "Abs."

2. Several problems come up regularly in evaluating "Adequacy of K." Each will be considered below.

   a. Omissions. Omissions may be either deliberate or inadvertent. For evaluating whether a writer has adequately expressed a K by way of inference, it is useful to distinguish between functional redundancy and non-functional redundancy. The former is needed for clarity of the message, but the latter is not.

   b. Substitutions or paraphrases. In evaluating substitutions and paraphrases, the main criterion again is whether the message is clear and essentially unchanged. To facilitate making these decisions, a worksheet is attached, with recurring substitutions evaluated. Add to the list from time to time, as decisions are made on other recurring substitutions.

APPENDIX A - Continued

c. Added information. Added information will vary on two scales, relatedness and correctness. If basically unrelated to the K's, it is disregarded when evaluating K's. Similarly, if related and correct, it does not influence the evaluation. Added information that is related to a K, and tends to obscure, distort, or contradict the original message, should result in a lowered "adequacy" evaluation.

d. Duplicated K's. If a K is expressed in two places, and the two versions are not contradictory, record the judged adequacy of the better version. If the two versions of a K are contradictory, grade down according to the amount of confusion or distortion introduced. "Low" will usually be assigned. (For example, Protocol 24 indicates that the chemical is liquid in T-8, and that it is a powder in T-9.)

e. Misspellings. Misspellings are not penalized, unless they obscure meaning.

Note: Some will write "alumin," rather than "aluminum" or "alumina." Mark down for this.

f. Unclear or inappropriate antecedents for reference words. This problem occurs frequently. Depending on the resulting misinformation or lack of clarity, check "Med" or "Low."

# APPENDIX B

## Materials for Scoring Interviews

### I. Interview Rating Sheet

| | Narration/ General Convers'n | Technical/ Academic/ Vocational | Overall |
|---|---|---|---|
| Accent | | | |
| Grammar | | | |
| Vocabulary | | | |
| Fluency | | | |
| Communic. | | | |

Listener: _____

List. Date: _____

Ss Name _____

Ss TOEFL No.-----

Location: Tokyo
(Circle one.) Lima
Santiago

| | |
|---|---|
| Accent | |
| Grammar | |
| Vocabulary (Nar., Gen c.) | Vocabulary (Tech, A, V) |
| Fluency (N, Gc) | Fluency (T, A, V) |
| Communic. (N, Gc) | Communic. (T, A, V) |
| General Comments: | |

APPENDIX B - Continued

II. Proficiency Descriptions

A. Accent

1. Pronunciation is generally unintelligible.

2. Frequent gross errors and a very heavy accent make understanding difficult. Meaning of some portions of interview lost or greatly obscured.

3. Marked "foreign accent" requires concentrated listening, and leads to some lack of clarity in the message and apparent errors in grammar or vocabulary.

4. A noticeable accent and occasional mispronunciations are present but do not interfere with understanding.

5. No conspicuous mispronunciations, but would not be taken for a native speaker.

6. Native pronunciation, with no trace of foreign accent.

B. Grammar

1. Shows virtually no command of grammar except for stock phrases.

2. High error rate or use of only very few grammatical patterns, frequently limiting or impairing communication.

3. Error rate or limited use of grammatical patterns causes occasional misunderstanding. Control of basic patterns is shown.

4. Errors generally limited to those having little effect on understanding. Subject uses longer, more complex sentences, and makes effective use of expressions such as "would," "should" "as soon as," etc.

5. No systematic errors that influence understanding. Subject demonstrates command of nearly the same range of grammatical usages that would be expected of an American college student in the same interview.

6. Subject's grammar is indistinguishable from that of a native English speaker at the same level of education.

APPENDIX B - Continued

C. Vocabulary

| General | Academic/Professional |
|---|---|
| 1. Vocabulary inadequate for even the simplest conversation or narration. | 1. Academic or professional vocabulary limited to a handful of words, insufficient for conversation. |
| 2. Vocabulary is very limited even in such basic topics as family, home, travel, and time. Narrative (for Latin American subjects) is severely limited because of vocabulary problems. | 2. Discussion in professional or academic areas is severely limited. |
| 3. Vocabulary is adequate for only a limited range of general topics. General conversation or narrative is weakened by inaccuracies and limitations in vocabulary. | 3. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional or academic topics. |
| 4. Vocabulary permits discussion of a wide range of general topics, with circumlocutions. Narrative is not noticeably ably curtailed by vocabulary. | 4. Vocabulary is adequate for a moderate range and depth of discussion in academic or professional areas. |
| 5. Considerable range and depth of vocabulary is demonstrated in general conversation or narration. The subject would not be mistaken for a native English speaker, but vocabulary limitations are minor. | 5. Vocabulary permits extensive discussion of subject's professional area. Though the subject is not a "native speaker," his/her vocabulary presents little if any barrier to communication. |
| 6. Vocabulary is apparently as accurate and extensive as that of a native speaker. | 6. Vocabulary is apparently as accurate and extensive as that of a native speaker. |

Appendix B - Continued

D. Fluency (General and Technical Conversation)

1. Speech is so halting and fragmentary that communication is virtually impossible.

2. Speech is very slow and uneven except for short or routine sentences.

3. There are smooth portions but speech may become hesitant, jerky, or sprinkled with "ah...," "er...," etc.

4. Speech is generally smooth with some pauses for rephrasing, groping for words, etc.

5. Speech is effortless and smooth but perceptibly non-native.

6. Speech on all topics is as effortless and smooth as that of a native speaker.

E. Communication (General Conversation)

1. Subject's ability to communicate even very basic information is virtually nil. This is evident both in his responses, per se, and in his apparent inability to understand the interviewer's statements.

2. Interviewer must speak slowly and simply to subject. Subject's communicating of his/her own ideas is limited to short, simple sentences, often too fragmentary to be understood.

3. Subject understands a good deal of what is said to him with some simplification and rephrasing on part of speaker. In his/her own conversation, subject has some difficulty in getting the message across, relying on awkward circumlocutions, rephrasing, and occasional help from the interviewer.

4. Subject understands quite well normal speech both in general and technical areas. With occasional repetition and rephrasing, the subject can continue a conversation or a narrative; however, lack of control of some speech patterns and vocabulary limitations prohibit very sophisticated discussion.

5. Subject communicates both simple and complex ideas effectively. He/she may occasionally resort to circumlocutions or express self in a somewhat "foreign" way (e.g., by using nonstandard word-order), but these deficiencies impose essentially no limitations on what he/she discusses, or on the clarity of what he/she says.

6. The subject's ability to communicate is indistinguishable from what would be expected from an American counterpart.

Appendix B - Continued

3. Training Tapes

Training Tape (Cam)

Accent - Subject spoke with decided Spanish accent, often making such mispronunciations as "dere" for there and "dey" for they: This factor did not interfere with understanding.

Grammar - She used a variety of tenses, both simple and compound ("you promised me that you would pay back...") with a high degree of success. Some awkwardness with such phrases as: "the fox, he gets mad" or "in few seconds."

Vocabulary - The narrative part of the interview demonstrated a large vocabulary (such terms as "rural," "tropical," and an idiomatic, "Take it easy"). Subject was able to elaborate rather than just answer a question.

Fluency - Some groupings and hesitations interrupted the smoothness of the speech.

Communication - Speaker had no problem understanding interviewer. Her own narrative attempt was well developed and cohesive so that she got a rather complex message across.

| | General Conversation | Academic Conversation | Overall |
|---|---|---|---|
| Accent | X | X | 4 |
| Grammar | X | X | 5- |
| Vocabulary | 5 | 4+ | 5 |
| Fluency | 4+ | 4 | 4+ |
| Communication | 5 | 5 | 5 |

Appendix B - Continued

Training Tape (Shibuya)

Accent - Marked Japanese accent which demands a concentrated listening. Mispronunciations common ("har douter" for her daughter, "bisit" for visit, "perfrectry" for perfectly) and tends to drop final consonants of some words.

Grammar - Uses present and simple past tenses while making some errors in subject-verb agreement and present perfect tense. Tends to answer questions in fragments whenever possible.

Vocabulary - Subject has a certain facility with his limited vocabulary, but avoids discussing any complex general or academic matters.

Fluency - Frequently hesitant, uses fragments.

Communication - Subject appears reluctant or unable to complicate his simple conversation with the interviewer. Interviewer must repeat or rephrase some simple questions (e.g., "Tell me about the place you live"). Initially, subject confused "export" with "transport" but then corrected himself.

| | General Conversation | Academic Conversation | Overall |
|---|---|---|---|
| Accent | X | X | 3 |
| Grammar | X | X | 3 |
| Vocabulary | 3 | 3 | 3 |
| Fluency | 3 | 3 | 3 |
| Communication | 3 | 3 | 3 |

Appendix B - Continued

Training Tape (Vaca)

Accent - Accent extremely thick. Listener must strain to understand and still fails to catch some phrases. A typical pronunciation example: "did" for died.

Grammar - Subject omits articles, often uses present tense when past is desired (although can use some simple past tenses) and lapses into his native tongue quite readily. He speaks in fragments.

Vocabulary - Range of words extremely limited. Could use words like "brother," "chair," and "work," but faltered on more technical words.

Fluency - Subject's speech was hesitant, stammering with much fumbling for vocabulary and much repetition; in a word, non-fluent.

Communication - In spite of grammatical and vocabulary problems, subject was able to understand and get across some information in English during the interview.

| | General Conversation | Academic Conversation | Overall |
|---|---|---|---|
| Accent | X | X | 1+ |
| Grammar | X | X | 2- |
| Vocabulary | 2- | 2- | 2- |
| Fluency | 1 | 1 | 1 |
| Communication | 2- | 1+ | 2- |

Appendix B - Continued

Training Tape (Tokyo 42)

Accent- Very slight. Although it would be impossible to say which country the subject is from, "an alert listener" would detect that English is not her native language.

Grammar- Very few mistakes. Only prominent error occurred in the subject's description of her favorite teacher ("The thing I liked is she gives us homework..."). Several minor mistakes. Interviewee used many and varied complex constructions.

Vocabulary- Extensive, with good knowledge of idioms and common usage ("pretty large yard," "on top of a hill," "something like that," "manage to accomodate").

Fluency- Pauses only as much as a native speaker might.

Communication- Understands virtually everything said to her, and her own statements are complete and clear.

| | General Conversation | Academic Conversation | Overall |
|---|---|---|---|
| Accent | X | X | 5+ |
| Grammar | X | X | 6- |
| Vocabulary | 6 | 6- | 6 |
| Fluency | 6 | 6 | 6 |
| Communication | 6 | 6 | 6 |