DOCUMENT RESUME

ED 323 786 FL 018 843

AUTHOR	Stansfield, Charles W.; And Others
TITLE	Listening Summary Translation Exam (LSTE)-Spanish. Final Project Report. Revised.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
PUB DATE	31 May 90
NOTE	120p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	English; *Language Tests; *Listening Comprehension; Program Evaluation; *Spanish; *Standardized Tests; Tape Recordings; *Test Construction; Test Reliability; Test Validity; *Translation
IDENTIFIERS	Federal Bureau of Investigation; *Listening Summary Translation Exam Spanish

ABSTRACT

The Listening Summary Translation Exam (LSTE)-Spanish is designed to assess the ability to comprehend and summarize in written English recorded conversations spoken in Spanish. Language and topics of the exam are representative of conversations the Federal Bureau of Investigation routinely monitors. This test consists of two subtests, one containing 40 multiple-choice items based on eight to nine recorded conversations, and the other requiring written summaries of three recorded conversations. The LSTE-Spanish exists in two forms parallel in content, item difficulty, format, and length. The final report on the test's development is presented in four sections. The first provides a general description of the test, and the second describes the development of the test's two pilot forms. In the third section, the results of the trials and piloting of the pilot forms are presented, with descriptions of the resulting revisions. The fourth section presents the procedures and results of a study to validate the LSTE's final operational forms. Substantial appended materials include selected pages of the test booklets, guidelines and forms, and pilot testing, validation, and feedback results. (MSE)

LISTENING SUMMARY TRANSLATION EXAM (LSTE) - SPANISH

Final Project Report
Revised

Charles W. Stansfield
Mary Lee Scott
Dorry Mann Kenyon

Center for Applied Linguistics
1118 22nd Street, N.W.
Washington, D.C. 20037

May 31, 1990

LISTENING SUMMARY TRANSLATION EXAM (LSTE) - SPANISH

Final Project Report
Revised

Charles W. Stansfield
Mary Lee Scott
Dorry Mann Kenyon

Center for Applied Linguistics
1118 22nd Street, N.W.
Washington, D.C. 20037

May 31, 1990

Table of Contents

Appendices

II. SPANISH LISTENING SUMMARY TRANSLATION EXAM

This part of the final rep rt is divided into four sections. The first section provides a general description of the operational version of the Spanish Listening Summary Translation Exam (LSTE-Spanish). The second describes the development of the two pilot forms of the test. In the third section, the results of the trialing and pilcting of the pilot forms are presented, together with a description of revisions made in the forms. The fourth section presents the procedures and results of a study to validate the final operational forms of the LSTE-Spanish.

1. General Description

The Spanish Listening Summary Translation Exam (LSTE-Spanish) is designed to assess the ability to comprehend and summarize in written English recorded conversations spoken in Spanish. The language and topics of the exam are representative of the conversations which the FBI routinely monitors.

The LSTE-Spanish consists of two subtests. The first contains 40 multiple choice items based on eight to nine recorded conversations. This subtest is referred to in this part of the report as the Multiple Choice section. The second subtest requires examinees to write summaries of three recorded conversations. This subtest is referred to as the Summary section. A separate test booklet for each section contains instructions, example items, and test items. A master tape for each section contains the general introduction to the exam,[1] instructions, example items, and recorded conversations. The LSTE-Spanish exists in two forms that are generally parallel in content, item difficulty, format, and length.

1.1 Multiple Choice Section

This section of the report describes the format, and test taking and scoring procedures for the Multiple Choice section of the LSTE-Spanish.

[1] Examinees are informed that they will hear brief conversations involving two people, and that the age, sex, and regional accent of the speakers will vary.. The following disclaimer is then offered:

> The language and topics are representative of the conversations which the Bureau routinely monitors. However, the names and characters are entirely fictitious and any resemblance to actual individuals is purely coincidental. The opinions expressed do not reflect those held by the Bureau.

1.1.1 Format

There are 40 items in the Multiple Choice section, based on eight to nine recorded conversations. These conversations simulate exchanges regarding drug deals, fraud, terrorism, illegal immigration, and foreign counter intelligence. Because they are unscripted, the conversations manifest all of the characteristics of natural speech, including hesitations, false starts, repetitions, interruptions, overlapping of speakers, misunderstandings, requests for clarification, etc.

The test items vary in purpose: some of them assess comprehension of specific details such as dates, times, locations, etc., while others require the examinee to infer the relationship of the speakers, their emotional reactions to the messages conveyed, and possible actions to follow from the conversations.

A test booklet contains instructions, example items, explanations, and the test items themselves. Appendix A contains selected portions of a test booklet for the Multiple Choice section, including the cover page, instructions, and example items.

1.1.2 Test Taking

Each examinee receives a Multiple Choice section test booklet, a machine scoreable answer sheet, and two no. 2 pencils. Examinees listen to the exam instructions on the tape, and read along in their test booklets when instructed to do so.

Examinees are informed that they will hear a series of conversations, some of which are related to each other. In this section, each conversation is presented only once. Examinees are given a block of time before hearing a given conversation to scan the questions and options pertaining to that particular conversation.[2] By scanning the items before hearing the conversation, they have an idea of what type of information to listen for.

As they listen to the conversation, examinees may read the items again and mark their choices in the test booklet. They are cautioned not to be distracted by slang or phrases which are unfamiliar to them. Instead, they are to concentrate on extracting only the information needed to answer the questions.

After listening to the conversation, examinees are given a another block of time to review their choices and transfer their answers to the machine scoreable answer sheet.[3]

This section lasts approximately 35 minutes.

1.1.3 Scoring Procedures

Examinees record their responses to the Multiple Choice section of the LSTE-Spanish on answer sheets which are scored by machine. The score on this section is the number of answers correct. The maximum possible score is 40.

1.2 Summary Section

This section describes the format, and test taking and scoring procedures for the Summary section of the LSTE-Spanish.

1.2.1 Format

In the Summary section, examinees are required to summarize three conversations, which increase in length (from approximately one to three minutes) and in sophistication of vocabulary. The conversations are similar to those in the Multiple Choice

[2] There are four to six items for every conversation. Examinees are given from four to six seconds to scan each item, depending on the length and complexity of the item.

[3] They are given from 12 to 18 seconds per item, depending again on the length and complexity of the item.

section. However, in this section examinees hear each conversation twice, and they are permitted to take notes on the content of the conversation.

The Summary section test booklet contains instructions, space for taking notes and writing a summary of an example conversation, observations regarding the example summary, and space for taking notes and writing summaries of the remaining conversations. (Appendix B contains selected portions of the test booklet for a Summary section, including the cover page, instructions, an example summary, and an analysis of the example summary.)

1.2.2 Test Taking

In this section, examinees hear each conversation twice. They take notes as they listen to the conversation, and then write a summary in English using the information in their notes.[4]

Examinees are told what kind of information should be present in an effective summary, including the overall topic of the conversation, and supporting details including names, dates, times, places, or amounts. As conversations vary in the amount of concrete information they contain, examinees are cautioned to make sure they identify the general topic and primary supporting points of more abstract conversations. They are instructed to include as much detail as possible in the summary. However, they are to include only information they have gleaned from the conversation, and not to add any of their own assumptions or inferences.

The duration of this section is approximately 45 minutes.

[4] Examinees are given from three to ten minutes to write summaries of the conversations; the amount of time allotted depends on the length of the conversation. Before beginning a particular summary, examinees are informed of how much time they will be given. They are also advised when there is one minute remaining to complete the summary.

1.2.3 Scoring Procedures

Examinees receive two scores for this section: one for Accuracy and the other for written Expression. Both are assessed by a trained rater.

Accuracy is scored by the rater through the use of a checklist containing the main topic, key and supporting points in the conversation. An example of a checklist, based on the example conversation, is provided in Appendix C. As the rater reads a summary, he or she checks off those items on the list which the examinee has reported accurately; two points are awarded for the main topic, one point is awarded for each key and supporting point. Although the wording of the summary does not have to match exactly that of the checklist, it is important that the information be provided in the appropriate context. Because the content of the conversation is broken down into items of information on the checklist, an examinee can receive credit for each item that is accurately reported, even if other items are omitted or misunderstood. The total Accuracy score is the sum of the points awarded for each of the three conversations. The maximum number of points for Accuracy on Form 1 of the LSTE-Spanish is 74; on Form 2 it is 71.

Expression is scored by the rater through an evaluation of the written summary in terms of the correctness of grammar, spelling, punctuation, and syntax, and the effectiveness of vocabulary and organization it displays. This evaluation proceeds according to the Expression Scoring Guidelines (see Appendix D). For each of the three summaries, the examinee is awarded either a Deficient (= 1 point), Functional (= 2 points), or Competent (= 3 points). The Final Expression Rating is the average of the Expression scores on the three summaries. Once the average is computed, a final rating is awarded as follows:

Average Expression Score	Final Expression Rating
1, 1.33	Deficient
1.67, 2.00, 2.33	Functional
2.67, 3.00	Competent

The Accuracy and Expression scores on the LSTE-Spanish Summary section are always kept separate. However, a total score (TOT) for Accuracy on the LSTE-Spanish is awarded by adding the raw score on the Multiple Choice section and the Accuracy score on the Summary section together. The maximum total Accuracy score obtainable on Form 1 of the LSTE-Spanish is 114; on Form 2 it is 111.

Raw scores for Accuracy on the LSTE-Spanish (for section scores or total scores) can be converted to a Final Accuracy Rating (ranging from No Ability to Superior) through the use of the Final Accuracy Rating Conversion Table (Appendix E). The development Final Accuracy Rating Conversion Table is described in section 4 of this part of the report. An interpretation of the Final Accuracy Rating is provided for administrative purposes, describing typical performance on a summary writing task at each level on the scale (see Appendix R).

In order to pass the LSTE-Spanish, CAL proposes that an examinee must obtain a Final Accuracy Rating of at least Functional and a Final Expression Rating of at least Functional.

1.3 Use of Multiple Choice Section Score in Screening

The Multiple Choice section may be used to screen out individuals for whom the Summary section of the exam would be inappropriate; that is, examinees who would not be likely to achieve a Final Accuracy Rating of Functional or above. Through statistical analyses (described in section 4), we have determined that the raw score cut-off on the Multiple Choice section should be 19 for Form 1 and 16 for Form 2. Examinees scoring at or below these scores need not take the Summary section. If they have already taken this section, it need not be scored.

2. Development of the LSTE-Spanish

This section describes how the two pilot forms of the LSTE-Spanish were developed. The method of preparing and recording the simulated conversations, the preparation of examination materials, and development of the pilot study scoring methods are discussed.

2.1 Conversations

CAL had originally proposed using taped conversations from adjudicated cases provided by the FBI as the basis for the items on the LSTE-Spanish. However, the FBI later informed CAL that it would not be possible to use actual tapes. Instead, it was necessary for CAL to develop and record original conversations.

In preparation for the creation of conversations, we conducted an informal analysis of adjudicated tapes provided by the FBI in order to identify the general characteristics of the conversations typically monitored by the Bureau. The analysis included identification of frequent topics, tone, and use of nicknames, colloquial expressions, and code words. We then prepared a summary of the general characteristics we discovered. In addition, CAL consultants developed a number of brief scenarios outlining the gist of conversations to be used for the LSTE-Spanish.

CAL staff and consultants met with FBI staff to discuss the general characteristics of monitored conversations, the scenarios which had been developed to that point, and the exam format and scoring. As a result of this meeting, the original summary of characteristics was revised and expanded with information obtained from FBI staff (see Appendix F). Additional scenarios were subsequently developed so that conversations relating to drug deals, fraud, illegal immigration, terrorism, and foreign counter intelligence would be represented on the exam.

Once the scenari-s were developed, CAL brought in male and female native Spanish-speaking actors to improvise conversations based on the information in the scenarios. The actors varied in age and spoke a variety of dialects of Spanish including Puerto Rican, Dominican, Mexican, Peruvian, Bolivian, and Chicano. We briefed the actors before each taping session by reviewing the general characteristics of monitored conversations and playing several of the tapes of adjudicated conversations which the FBI had supplied. The actors were encouraged to speak naturally and use any slang, regionalism, or even vulgarities they felt would be appropriate in a given situation. An FBI staff member was present at recording sessions in order to provide feedback on the authenticity and acceptability of the conversations as they were being taped.

After reviewing the scenario for a given conversation, the actors agreed on code words and basic content, rehearsed the conversation briefly a few times face-to-face, then retired to different rooms and carried out the conversation by phone. The conversations were taped using a recording device attached to one of the phones, thus simulating as closely as possible conditions under which conversations are often recorded by the Bureau. A conversation was re-taped as many times as needed until it met the approval of CAL and FBI staff.

A total of 35 different conversations were taped over a number of recording sessions.

2.2 Exam Forms

CAL staff and consultants wrote multiple choice items based on a number of the recorded conversations. The items were designed to assess understanding of specific information and ability to make inferences based on the information presented in the conversations.[5]

[5] The items in the Multiple Choice section differed in this aspect from the instructions given in the Summary section, which cautioned the examinee not to insert his or her own inferences in writing the summary, but to report only the information present

Parallel forms of the LSTE-Spanish were constructed so as to ensure a similar distribution of the number of conversations (for each form, 12 in the Multiple Choice section and 3 in the Summary section), length of conversations, the sex of the speakers, and the number of multiple choice items which had been developed (60 items for Form 1 and 56 items for Form 2). After developing the answer key for the Multiple Choice section of each form, we made changes in the ordering of the options to ensure equal distribution of correct answers across the four choices A, B, C, and D. More conversations and items than would be needed on the final versions were prepared, so that only those which functioned most effectively could be retained. A summary of the content and format of the pilot versions of the LSTE-Spanish is located in Appendix G.

2.3 Exam Tapes

After organizing the conversations and items into parallel forms, we prepared scripts for the narration of each form. The scripts included a general description of the exam, instructions for filling out the machine scoreable answer sheet and test booklet, example items and explanations, multiple choice and summary item numbers, and instructions to the recording engineer for placement of the recorded conversations.

CAL worked with a professional recording studio, Lion and Fox, Inc., to edit and assemble the conversations into the two forms. The narration of the forms was recorded in the studio by a professional radio announcer. Subsequently, the narration and conversations for each form were merged on to a master tape. At this time the pauses before each conversation and between items were inserted. Cassette copies for use in the pilot study were made from the master tape.

CAL also prepared test booklets for each form of the LSTE-

in the conversation.

Spanish (as described in section 1 of this part of the report).

2.4 Pilot Test Scoring Procedures

Scoring procedures for the Multiple Choice section of the test, i.e., counting the number of correct answers, were straight-forward. For the Summary section, however, we wanted the scoring procedures to reflect, as closely as possible, the FBI/CAL translation skill level descriptions. Yet because the task of summarizing a spoken conversation differs from rendering a verbatim translation of a written document, we knew there would inevitably be differences in the two scoring systems. As the FBI/CAL descriptions refer to "minor" mistranslations or omissions (presumably in contrast to "major" or "substantive" mistranslations or omissions), we felt that it would be important to make this distinction with reference to the information to be included in the summaries. Consequently, we devised a plan to identify the "key" and "supporting" points in the Summary section conversations.

In order to do this, we wrote a summary of each of the conversations by listening to the conversation several times, stopping and re-playing the tape as often as needed in order to capture as much detail as possible. We then asked six FBI language specialists to read the summaries and underline the key points. Their responses were tabulated. Points identified by five out of the six experts were subsequently considered "key" points for the purpose of scoring. The remaining points were considered "supporting" points.

We then developed a checklist for each summary, similar to that described in section 1.3.2.

In addition to the checklist, we also developed a Summary Scoring Guide for use in evaluating Substantive Accuracy and Expression (in separate categories of Grammar, Spelling and Punctuation, Vocabulary, and Organization). The original guide was based on the FBI/CAL translation skill level descriptions.

It was refined in scoring the data from the pilot study described below. (A copy is located in Appendix H.) Using the guide, an examinee's performance was characterized as showing No Ability, or being Incompetent, Deficient, Functional, Competent, or Superior, and a numerical score was assigned in each category.

3. Trialing and Pilot Testing

This section describes two very important steps in test development: trialing, which may be considered a preliminary check using examinees before piloting, and piloting. The section discusses the results of the piloting and the subsequent revision process.

3.1 Trialing

CAL staff met twice with members of the FBI staff to evaluate the pilot versions of the exam forms. In addition, both forms were trialed informally on four CAL staff members with varying degrees of proficiency in Spanish.

In the pilot version, there was only one test booklet. Its format for the Multiple Choice section was modified somewhat after trialing and consultation with the FBI to ensure that all multiple choice items for a given conversation appeared on the same page. Moreover, additional blank lines for notes and writing the summaries were provided. We modified the directions to allow examinees to mark the answers to the multiple choice items in their test booklets while they were listening to the conversations, and then transfer them to the machine scoreable answer sheet in the time allotted after the conversation. Other suggested modifications were delayed pending the results of the pilot study.

3.2 Pilot Testing

In pilot testing, empirical data is gathered that will inform the test revision process. This section describes how the LSTE-Spanish was piloted.

3.2.1 Data Collection

The LSTE-Spanish was piloted on 31 university students (enrolled in varying levels of Spanish language instruction) and 15 FBI staff members from the Washington Field Office in late

July and early August of 1989. Twelve additional examinees, most of them native speakers of Spanish, took the exam. Not all examinees were able to take both forms: in total 49 took Form 1 and 37 took Form 2; of these, 23 took both forms.

The examinees in the pilot study completed a questionnaire that assessed their reactions to exam instructions, examples, and format. A copy of the questionnaire can be found in Appendix I.

The Multiple Choice section data was scored by machine. The written summaries from the Summary section were scored by two raters, a CAL staff member and a consultant, using the checklists and scoring guide described above.

3.2.2 Results

Table 3.1 presents a summary of the performance of all examinees included in the pilot study on the Multiple Choice section. Reliability estimates, calculated using Kuder-Richardson formula 20 (KR-20), are also displayed.

Table 3.1
Multiple Choice Section
Total Pilot Sample

Form	N	Mean[6]	%	Std. Dev.	KR-20
1	49	37.0	62	10.8	.91
2	37	40.8	73	9.7	.91

KR-20 yields an estimate of the internal consistency of the test items, i.e., a measure of the extent to which examinees perform consistently across the items within a test. As can be seen from Table 3.1, the reliability estimate was quite high for both forms. It was difficult to judge their comparability in terms of difficulty from this data, however, as there was a great deal of variation in the number and proficiency level of the

[6] Note there were 60 items on the pilot version of Form 1 and 56 on Form 2.

examinees who took each form.

In order to obtain an idea of the comparability of the exams, it was necessary to examine the mean scores of those individuals who took both forms. A summary of their performance is displayed in Table 3.2 below:

Table 3.2
Multiple Choice Section
Pilot Sample Subset

Form	N	Mean	%	Std. Dev.
1	23	43.9	73	8.6
2	23	42.4	76	6.1

Although the forms were fairly comparable in level of difficulty, Form 1 appeared to be slightly more difficult than Form 2.

The goal in raters' scoring of the Summary section of the pilot was to test the appropriateness of the scoring system and to select benchmark papers for rater training materials for the validation study (described in section 4.2.1). Thus, descriptive statistics were not calculated for this section.

Results of the examinee questionnaire are presented in Appendix J.

3.2.3 Revisions

We decided to include 40 multiple choice items on the final version of each form. In order to identify items which should be deleted or revised, we conducted an item analysis to discover which items appeared to be too easy or too difficult, or did not discriminate among examinees of higher and lower ability. Responses to the examinee questionnaire were also taken into consideration in revising the format of the exam.

A total of five conversations were deleted from the two forms of the LSTE-Spanish for a variety of reasons. In one case, the items corresponding to one of the conversations were too easy. Another conversation contained confusing items and didn't seem to flow naturally. Three conversations were deleted because the FBI felt the language used, although realistic, was too obscene to be included on an FBI test instrument. A number of phrases were edited out of two of the remaining conversations for a similar reason. Several remaining multiple choice test items were either deleted or revised.

As a result of suggestions from FBI staff members who took the test, as well as from other information gained during the pilot testing, several points were added to the general introduction to the exam.[7] The amount of time given to scan items before hearing the conversations was increased and varied according to the length and complexity of the item.[8] Similarly, the time allotted to review and transfer answers to the machine scoreable answer sheet was varied.[9] In addition, separate test booklets for the Multiple Choice and Summary sections were developed, so that the former could be discarded and the latter retained for scoring. Finally, in the Summary section, examinees were given the opportunity to actually write a summary based on their notes from the sample conversation, instead of merely reading an example of a summary.

In order to achieve a balance in length, content, number of items, sex of speakers, and topic, the remaining conversations were re-arranged between the two forms. As examinees in the

[7] Specifically, the introduction was expanded to include a statement to the effect that the characters portrayed were entirely fictitious and the opinions expressed did not reflect those of the FBI. Examinees were also cautioned not to be distracted by slang or phrases they didn't understand, but instead concentrate on extracting only the information needed to answer the questions.

[8] The time was varied from 4 to 6 seconds per item.

[9] The time was varied from 12 to 18 seconds per item.

pilot study had difficulty completing the longer summaries in the time allotted, selected portions of the final conversations were deleted to make them shorter and less complex, and examinees were given a little more time to write the summaries. An overview of the organization of the final version of the exam forms can be found in Appendix K.

4. Validation Study

The validation study for the LSTE-Spanish was an attempt to research to what extent the test is measuring summary writing skills. This section describes the study design, data collection procedures, test scoring procedures, results and discussion of the study.

4.1. Overview

The design of the validation study called for administering the LSTE-Spanish to FBI language specialists, agents, and other employees at field offices around the country. Efforts were made to include individuals of differing ability levels. In order to examine the validity of the LSTE, scores on other measures of language ability were obtained from employee files as available.

Both forms of the LSTE-Spanish were given in one sitting (about three hours in duration) at each of seven FBI field offices. The order of administration of the forms was counterbalanced to control for test practice effect. Thus, approximately half of the examinees took Form 1 first and the other half took Form 2 first.

4.1.1 Test Administration Instructions

CAL developed a set of test administration instructions. These include instructions to the test administrator regarding the following: 1) test security, 2) assembling test materials, 3) arranging for a testing site, 4) equipment, 5) administering the test (including timing of sections), and 6) procedures to follow after the test. Appendix L contains a copy of the administration instructions for the LSTE-Spanish.

4.1.2 Questionnaires

CAL developed two questionnaires for use in the validation study: 1) a self-assessment questionnaire on which an examinee was asked to estimate his or her ability to perform summary translation tasks, and 2) a questionnaire requesting examinee feedback on aspects of the format and content of the exam (similar to the pilot questionnaire). The self assessment questionnaire was administered prior to the LSTE-Spanish, and the exam feedback questionnaire following the LSTE-Spanish administration. A copy of the self assessment questionnaire is located in Appendix M and a copy of the exam feedback questionnaire in Appendix N.

4.1.3 Subjects

Testing materials, including test administration instructions, numbered test booklets, tapes, answer sheets, pencils, questionnaires, and test administrator report forms[10] were sent to the FBI field offices in Los Angeles, San Diego, Albuquerque, Phoenix, and El Paso on November 15, 1989. Similar sets of materials were sent to Houston[11] and Puerto Rico on November 17, 1989.[12] The test materials and answer sheets were returned to CAL within one to four weeks by the FBI field

[10] CAL developed this form for test administrators to note any irregularities that may occur with respect to test security, the test administration, or the condition of the test materials. We requested that the validation study test administrators complete and sign the form even if there were no irregularities. (See Appendix II-O for an example of this form.)

[11] Arrangements were made for members of the Houston Police Department (for whom Spanish OPI scores were available) to be tested along with the FBI employees at the Houston field office.

[12] A cover letter was sent with the materials to the contact person at each field office. In addition to thanking them for their assistance in carrying out the validation study, the letter emphasized the importance of test security, outlined the procedures for the test administration, noted the proposed administration date, and instructed them to return all materials to CAL immediately after the test administration. A checklist of the materials was enclosed with the cover letter. CAL retained a copy of the checklists and used them to verify that all of the materials were returned as requested.

offices.

Since most of the FBI examinees were already working in Spanish, there were no low level ability examinees among them. Thus, in an effort to ensure that the entire range of abilities of potential test takers in the operational program would be represented in the sample, the FBI and CAL arranged for 20[13] beginning Spanish language students at the CIA to take the LSTE-Spanish. CAL staff administered both forms of the LSTE-Spanish to these students on February 2, 1990.

Thus, a total of 67 examinees took the LSTE-Spanish in the validation study. Of this group, 20 (30%) were CIA Spanish language students, 15 (22%) were FBI Special Agents, 11 (16%) were FBI Language Specialists (or contract linguists, who do similar work), 11 (16%) were FBI support staff, and 10 (15%) were members of the Houston Police Department. The geographic distribution of the examinees was as follows: Houston, 16; Los Angeles, 7; Puerto Rico, 6; San Diego, 6; Phoenix, 5; Albuquerque 5; El Paso, 2; and Washington, D.C., 20.

4.2 Scoring

The Multiple Choice sections of the LSTE-Spanish forms were scored by machine, using answer keys based on the revised versions of the forms. The Summary sections were scored by trained raters. The development of rater training materials, and the rater training procedure is described in detail below.

4.2.1 Development of Rater Training Materials

As mentioned in section 3.2.2, the summaries from the pilot study were rated by two raters using checklists and the Summary Scoring Guide. In selecting benchmark summaries to illustrate the levels of the scoring guide, we identified those summaries to

[13] One of the CIA examinees left after taking only the Multiple Choice section of Form 1.

which the raters assigned the same score, or scores which differed only slightly. Where we could not find a summary on which the raters agreed at a particular level, we chose one on which they seemed to be in the closest agreement. In this way, benchmarks illustrating all levels (Incompetent through Superior) of Substantive Accuracy were selected for each of the summary conversation item per form. In addition, one benchmark summary was selected to illustrate different levels within each of the Expression subcategories (Grammar, Spelling and Punctuation, Vocabulary, and Organization). In addition, sets of practice summaries were chosen for all of the summary items. Each practice set contained five summaries of varying quality.

We prepared notebooks for the raters containing an example summary of each conversation,[14] summary checklists, the scoring guide, scoring sheets, and sets of benchmark and practice summaries for each summary item.

4.2.2 Procedure

On December 11, 1989, four people sent by the FBI, Olga Navarrete, Adriana Peroutka, Juan Mesas, and Jack Nixon, began training to rate the summaries from the validation study. Marijke Walker also participated in the first two days of the training. The training and scoring sessions were conducted by Charles Stansfield and Mary Lee Scott at CAL, and continued from December 11th through 14th, followed by one session on the 19th.[15]

Stansfield conducted the first two days of rater training. After giving a brief overview of the purpose and development of

[14] These were the summaries used to identify key and supporting points as discussed in section 2.5.2.

[15] Adriana Peroutka, Juan Mesas, and Jack Nixon completed the training on December 14th. Olga Navarrete was unable to attend that day, but continued her training on December 19th.

the LSTE, he explained the use of the Summary Scoring Guide, checklists, and scoring sheets. He then played the first summary conversation from Form 1, while the raters reviewed the example summary and checklist for that conversation as they listened. Subsequently, he instructed the raters to score the first of the benchmark examples for Substantive Accuracy, using the scoring guide. After a brief discussion, the group rated and discussed the other benchmark examples for that item.[16] In addition, they rated and discussed the first set of five practice summaries.

When it appeared that the raters had achieved a fairly high level of agreement on Substantive Accuracy ratings for the practice summaries, Stansfield introduced the Expression portion of the scoring guide. Benchmark examples of performance in each of the subcategories relating to Expression (Grammar, Spelling and Punctuation, Vocabulary, and Organization) were then reviewed and discussed. The final practice set of five summaries for the first summary on Form 1 were then rated for both Substantive Accuracy and Expression. This was again followed by a discussion. At this point, the raters scored the first summary on the rest of the papers that were obtained from the validation study sample.

The same procedure was followed for the other summary items from Form 1 and Form 2, except that as the raters became more familiar with the task, fewer practice summaries were scored. In an effort to control for any change in scoring standards by the raters as the become more familiar with the process, the order of the test forms scored was counterbalanced. For the first and third summaries on the test, Form 1 summaries were rated before Form 2, while for the second summary, Form 2 was scored before Form 1.

Each FBI rater scored the summaries of half of the examinees

[16] Each set of benchmarks was arranged in descending order of quality from "Superior" to "Incompetent."

who participated in the validation study. In addition, the summaries were distributed among the raters so that each rater scored one-third of the summaries rated by the three other raters. In this way, comparisons could be made among the raters in determining interrater reliability.

4.2.3 Revision of the Scoring Procedures

The results of the scoring session indicated some problems with the scoring system as originally designed. The first was a discrepancy between the raters' awarding of points for the Substantive Accuracy score and their tallying of points from the Accuracy checklist. Theoretically, a high correlation in one should ensure a high correlation in the other. However, although the raters agreed highly on the tallying of points in the checklist, they did not on the awarding of a Substantive Accuracy score. This is shown in Table 4.1.

Table 4.1
Interrater Agreement for the
Awarding of Substantive Accuracy Scores and the
Tallying of Accuracy Checklist Points
by Form and Summary

Form/Sum	Substantive Accuracy Score	Total Checklist Points
F1/S1	.81	.91
S2	.61	.88
S3	.79	.92
F2/S1	.59	.86
S2	.75	.93
S3	.64	.85

This indicated that the raters were having a much harder time using the Substantive Accuracy scoring system than the checklist. Upon further investigation, it appeared that one of the problems was the interpretation of whether the examinee got the main topic of the conversation or not. There were two problems with this in the raters' using the Substantive Accuracy scoring guide. First, the guide put a ceiling on the highest score (Deficient) that could be awarded if the main topic was not clearly represented in the written summary. Thus, when two raters disagreed about whether the main topic was there or not, their Substantive Accuracy scores could be wildly divergent. Second, although the guide intimated a maximum score of Deficient if the main topic was not clearly present, even when (according to the checklist) raters agreed the main topic was not present, some on some occasions would still award a score higher than Deficient if most of the key and supporting points were present, while some on some occasions would not go above the Deficient category. This also lead to wildly divergent ratings on a single summary.

In light of the above, it was decided that for the validation study and the operational program, the Accuracy score would be based on the sumtotal of points present in the written summary. Each key and supporting point would receive one point while the main topic would receive two. As shown in section 4.3.2, this procedure led to a very high reliability (even though the original checklist ratings were used; i.e., the 47 examinees rated by the group of four FBI raters were not rescored for Accuracy).

After the group of 19 from the CIA was tested in February, their summaries were scored for Accuracy by two raters, Charles Stansfield and Mary Lee Scott, using the checklist only.

The second problem with the original scoring system was with the Expression scores. As shown in Table 4.2, the raters had tremendous difficulty agreeing on the scores in the various Expression subcategories.

Table 4.2
Interrater Agreement for the
Expression Subcategory Scores
by Form and Summary

Form/Sum	Grammar	Spelling/ Punctuation	Vocabulary	Organization
F1/S1	.39	.05	.36	.15
S2	.19	.37	.41	.03
S3	.30	.38	.38	.49
F2/S1	.34	-.10	.34	.04
S2	.04	.15	-.03	.29
S3	.19	.28	.22	.12

In light of the poor correlations between the raters' scoring in these subcategories, it was decided that Expression be rated globally on a scale having three categories: Deficient, Functional, and Competent. It seemed that these distinctions most adequately served to characterize most of the summaries. In this system, Functional and Competent are passing scores for Expression, while Deficient is not acceptable.

For the rest of this discussion, then, only scores following the revised procedures, which are the operational procedures, were used. As Expression scores from the original raters could not be used, a subset of 30 exams (15 of each form) were rated using the new Expression scoring system. These results, together with all results from the validation study using the revised scoring system, are presented in the next section.

4.3 Results

The results of the validation study test administrations are presented in this section by subtest.

4.3.1 Multiple Choice Section Descriptive Statistics and Reliability

Table 4.3 below presents the results of the validation study administration of the Multiple Choice section of the two test forms. This section in the two forms is referred to here as MC1 and MC2.

Table 4.3
Descriptive Statistics for MC1 and MC2

Form	Mean	Std. Dev.	Minimum	Maximum
MC1	28.97	6.63	12	39
MC2	27.39	7.26	11	38

Table 4.3 indicates that MC2 may be slightly more difficult than MC1. The larger standard deviation for MC2 suggests that less competent examinees may have tended to score slightly lower and more competent examinees slightly higher on MC2 than they did on MC1. Still the differences are not great.

The mean of MC1 represents approximately 72% correct while the mean of MC2 represents approximately 68% correct. Thus, for the group as a whole, the tests tended to be slightly easy, since we would expect a mean around 62.5% on a multiple choice test of optimal difficulty if the sample fully and equally represented the total range of abilities. It may be noted that the lowest score was just above what could be expected by chance (10 correct), but none of the subjects received a perfect score of 40. Thus, while a good range was represented in the sample, the sample contained more high ability students than low ability students as measured by the Multiple Choice section of the test. It should be remembered that the Multiple Choice section was intended to be used as a screen; i.e., to identify candidates who need not take the Summary section of the test. Thus, adequate performance on the Multiple Choice section would be a prerequisite to taking the rest of the test. If the total test is appropriate for the total sample, then it is not surprising that the Multiple Choice section would be slightly easy for the total sample. The high scores for this sample on the Multiple Choice section, then, are consistent with its intended use.

Table 4.4 presents the KR-20 reliability estimates for the two forms of the Multiple Choice section based on the validation study sample.

Table 4.4
KR-20 Reliability for MC1 and MC2

Form	KR-20
MC1	.86
MC2	.88

The reliability of the Multiple Choice sections of the LSTE-Spanish for both forms is relatively high and indicates that either form can be used with confidence on a population similar to that of the validation study.

A second indication of the reliability of the subtest is the consistency of performance of the group of 66 subjects on the two forms. Often called a coefficient of equivalence or parallel form reliability, it is the Pearson Product Moment correlation between subjects' performance on the two different forms. In the case of the LSTE Multiple Choice section, the coefficient of equivalence is .86. This is adequately high and represents the extent to which one form can be substituted for the other.

4.3.2 Summary Section Descriptive Statistics and Reliability

Table 4.5 below presents the results of the validation study administration of the Summary section of the two test forms scored for accuracy of the information contained in the summary using the Accuracy checklists. These Accuracy scores, representing the mean of the two raters' total points awarded, are referred to here as ACC1 and ACC2.

Table 4.5
Descriptive Statistics for ACC1 and ACC2

Form	Mean	Std. Dev.	Minimum	Maximum
ACC1	43.41	17.95	0	70.50
ACC2	36.46	17.18	0	66.50

Because there was a different number of total possible points on the two forms, the raw score means cannot be directly compared. However, the mean score on ACC1 represents approximately 59% of the possible total of 74 points and the mean score of ACC2 represents approximately 51% of the possible total of 71 points. Thus, ACC2 is more difficult than ACC1. (Note: this difference is accounted for in the scale score conversion described below.) However, as optimal difficulty for a test of this sort for the population should be 50% correct, the difficulty level of the test appears quite appropriate for the sample in the validation study. This gives us confidence that measurement statistics and the regression equations run to make the scaled score conversion table (see section 5) are accurate.

Since the Summary section is scored by raters using a checklist, the most important reliability to examine is the extent to which the raters agree when applying the checklist to the summary translations independently. This is a form of interrater agreement, and it would be expected to be quite high, since the checklist has been designed to be more objective than other rating schemes.

To determine the interrater reliability of using the checklist, a Generalizability study (G study) was performed on the rating data. A G study is a statistical technique in which the contributions of various factors (facets) to the total variance of the test scores are estimated. In our study, we wanted to estimate how much of error variance was being contributed by the raters; i.e., the effect of potential differences among rater ability. There were six raters involved in the study. All tests received two ratings; however not all raters rated all tests. Thus, a nested G study design was used. Following the G study, a Decision study (D study) was undertaken. A D study uses the findings of a G study to make decisions about efficient measurement procedures. In our study, we wanted to estimate the generalizability coefficient for the operational program using either one or two raters. A G coefficient is an estimate of reliability and is the ratio of the variance of the objects of measurement (in this case persons) over that variance plus error variance. The results of both studies are presented in Table 4.6.

Table 4.6
Results of the G and D Studies
on the Accuracy Scores for the Summary Section

Source of Variance	Variance	% of Total Variance
ACC1		
persons	319.21	98%
error	5.91	2%
ACC2		
persons	292.15	98%
error	5.87	2%

Form	G-Coefficient	
	1 rater	2 raters
ACC1	.98	.99
ACC2	.98	.99

Table 4.6 indicates that the amount and percentage of score variance due to error (the interaction between raters and persons; i.e., the inconsistency of the raters) is very small. The results of the D study indicate that the average reliability of a single rater in scoring using the checklist is very high

(.98) and that using two raters in the operational procedure is probably not necessary.

As with the Multiple Choice section, another indicator of the reliability of the test is to look at the consistency of performance of the group of 66 subjects on the two forms. For the Summary section, the coefficient of equivalence is .93, even higher than that of the Multiple Choice section. This is a good indicator that the two forms can be used interchangeably. However, only scaled scores (see section 5), that take into account differences in number of items and difficulty, can be compared. Raw score performance cannot be compared.

The Total Accuracy score is the sum of the scores on the Multiple Choice raw score and the Summary section Accuracy score. Although it is not possible to give an empirical reliability estimate for this total score, since the reliability of the Multiple Choice section and the Summary section are calculated in different manners, it is possible to give a coefficient of equivalence for Total Accuracy scores on the two test forms. This is the correlation coefficient between the Total Accuracy scores of the 66 subjects on the two forms and is identical to parallel form reliability. This coefficient .95 and is very high, indicating that the reliability of the composite score is most likely quite high. Given such a high coefficient, conservatively speaking, the reliability of the total score would not be lower than .86 for Form 1 or .88 for Form 2; i.e., their respective KR-20 reliabilities for the Multiple Choice sections. However, the reliability of the Total Accuracy score is likely much higher.

Each examinee also receives an Expression score on the Summary section of the test. To estimate the interrater reliability in awarding an Expression score, a subset of 15 examinees on Summary section Form 1 and 15 examinees on Form 2 were rated by 2 different raters. The extent of agreement on their Expression scores (the average score awarded over the three summaries) was .84. Since the range of scores was only between 1 and 3, this correlation may be misleadingly low. When the category placement is considered (i.e., whether the examinee was Deficient, Functional or Competent), of the 30 cases, there was agreement on three of the four subjects who were awarded a Deficient. For the other 26 cases, there were six cases in which one rater awarded the examinee a Competent while the other gave the examinee a Functional (both passing ratings). Thus, for the 30 cases, 23 (77%) were in complete agreement, and 29 cases (97%) would have been correctly assigned "pass" or "no pass" ratings. It appears that the Expression scale can be applied with a very high degree of consistency.

4.4 Examining the Validity of the LSTE-Spanish

According to the Standards for Educational and Psychological Testing (American Psychological Association, 1985), test validity refers to "the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (p. 9). Validity is demonstrated by an accumulation of evidence that supports the claim of validity for a particular test. Some of this evidence is empirical. Other evidence may be qualitative, in that it deals with the content of the test, or it may be theoretical, in that it deals with a theory about the nature of the trait being measured by the test. In the case of the LSTE-Spanish, the central validity concern is the claim that the test is a measure of the ability to summarize in standard written English the content of a conversation in Spanish.

Traditionally, three types of validity are usually identified according to how the evidence was gathered. These are content validity, criterion-related validity, and construct validity. Construct validity, which "focuses primarily on the test score as a measure of the psychological characteristic of interest" (APA, p. 9), may be understood to subsume the other two types; i.e., content and criterion-related validity are also evidences of the construct validity of a test. We turn first to a discussion of the content validity of the LSTE-Spanish.

4.4.1. Content Validity

Content validity is evidence that demonstrates the degree to which the sample of items, tasks or questions on a test are representative of the domain of content that could be tested. In the case of the LSTE-Spanish, evidence for its content validity is found in the tasks examinees are asked to perform to demonstrate their ability in listening summary writing. First, the Multiple Choice section checks their ability to understand conversations typical of those heard on-the-job. Clearly, without the ability to understand a conversation, there will not be the ability to summarize it. Second, the Summary section checks not only their understanding (the Accuracy score), but also their ability to convey their understanding in well written English (the Expression score). In this case, the task directly replicates what is called for on the job. It should be noted that there are two issues here--the accuracy of the information and the acceptability of the use of English in the summary. If the information in the summary is not correct, the summary is of no use to an investigation. On the other hand, if the information is correct but the expression is poor, then the summary could be discredited in a court of law.

The section describing the development of the taped telephone conversations highlights how close the conversations on the test simulate actual job-relevant conversations. The conversations on the test grew out of an analysis of actual taped conversations provided by the FBI. Furthermore, the test conversations were authenticated by FBI language specialists. Additional support for the content validity of the stimuli is provided by responses from the validation study subjects (agents and language specialists) to the exam feedback questionnaire they completed after taking the test (see Appendix P). On this questionnaire, 72% either agreed or strongly agreed with the statement "The conversations (in both parts A and B) were representative of the types of conversations I might encounter in my work."

At the same time, 59% percent of the subjects either agreed or strongly agreed with the statement "There was sufficient opportunity for me to demonstrate my ability to understand and summarize conversations spoken in Spanish." It may be that the 41% who disagreed with this statement did so because they felt unduly restricted by the time constraints of the testing situation: about half (49%) of the subjects felt the pauses for scanning the questions before the conversations were "too short" (51% felt they were "about right"), 36% felt the pauses for marking answers on the answer sheet were "too short" (62% felt they were "about right" and 2% felt they were "too long"), and 32% felt the pauses for writing the summaries were "too short" (68% felt they were "about right").

In interpreting these results, it is important to note that approximately 26% of those who took the LSTE-Spanish in the validation study had received scores of less than 2 on the Spanish OPI (see section 4.4.3 below); these subjects may have understandably felt pressured by the exam time constraints. On the other hand, those subjects whose proficiency was very high may have felt they didn't have sufficient time to incorporate all of the information they had recorded in their notes into their summaries. Though this may have been the case, these examinees would still have had ample opportunity to demonstrate their competence within existing time constraints.

4.4.2 Criterion-related Validity

Criterion-related validity is evidence that "demonstrates that test scores are systematically related to one or more outcome criteria" (APA, p. 11). For example, if there were an extant valid and reliable test of listening summary writing ability, then it would be important to see how scores on the LSTE-Spanish and scores on that test compare. Unfortunately, there is no other test that measures the same construct of listening summary writing ability that could be used as a criterion variable. In this case, a fuller discussion of evidence for the construct validity of the test is important. Such a discussion can be obtained by considering the convergent/divergent nature of the correlations with other measures theoretically related to the construct of interest. In such a discussion, an expected correlation of the test with each variable is analyzed and discussed. Some criteria will be expected to correlate highly with the test whose validity is being examined, while other criteria will be expected to correlate only moderately. Still other criteria might not be expected to correlate at all, or even to correlate negatively. We will make use of the convergent/divergent validity approach here in order to examine fully the construct validity of the LSTE-Spanish.

4.4.3 Convergent/Divergent Construct Validity

In an effort to provide evidence for the construct validity of the LSTE-Spanish, the following measures were also collected as part of the validation study:

1. A self-rating (SELF-ASMT). A self-assessment questionnaire that asked subjects to rate themselves on their summary writing ability was developed and administered to the subjects immediately preceding the administration of the first part of the LSTE-Spanish. A copy of this self-assessment is contained in Appendix M. 1 point was awarded for Limited, 2 for "Functional, 3 for Competent, and 4 for Superior. The self-assessment score was the sum of the scores on the three questions. The lowest possible score was 3; the highest, 12.

2. A Spanish OPI score (SPANSPK). An oral proficiency interview (OPI) score for Spanish was collected on each subject. Although this could not serve as a criterion variable, it is relevant to summary writing ability. Speaking proficiency assumes and is highly correlated with listening proficiency. Correlations between the

two skills on the ILR scale typically reach or exceed .90. Thus, on a theoretical basis, it was decided that the OPI score could be used to provide additional evidence of criterion-related validity. For all ILR scores in this study, the following conversion was used for purposes of empirical analyses:

ILR Score	Numerical Score
0+	0.8
1	1.0
1+	1.8
2	2.0
2+	2.8
3	3.0
3+	3.8
4	4.0
4+	4.8
5	5.0

3. <u>Other test scores</u>. Other scores that measure possibly related constructs were collected as possible. None of these scores could be collected for all the subjects, however. These scores, the number of subjects for which they were collected, and their descriptive statistics are given below.

Measure	N	Mean	Std. Dev.	Minimum	Maximum
DLPTLIST	30	52.70	4.95	39	60
DLPTREAD	30	53.23	6.31	30	60
SPANLIST	25	1.8	1.41	0.8	5.0
ENGSPK	18	4.2	0.58	3.0	5.0
SPENTRAN	18	3.36	1.00	1.8	4.8
ENSPTRAN	18	3.22	0.70	1.8	4.0

Key

DLPTLIST	The listening section of the Defense Language Institute Placement Test. Maximum possible score = 60.
DLPTREAD	The reading section of the Defense Language Institute Proficiency Test. Maximum possible score = 60.
SPANLIST	A Spanish listening score based on an ILR interview.
ENGSPK	An OPI score for English.
SPENTRAN	An ILR score on the current FBI Spanish into English verbatim translation exam.
ENSPTRAN	An ILR score on the current FBI English into Spanish verbatim translation exam.

4.4.3.1 Convergent Validity

As mentioned above, one method of establishing construct validity is to examine the divergence and convergence in the correlation of measures of traits that should theoretically be related or unrelated to the test. Correlations between the Total Accuracy scores on the LSTE-Spanish and its two sections with all the measures described above are presented in Table 4.7 below. The number of subjects involved in the correlation is also given, since not every subject had a score on every measure. (The numbers in parentheses represent the number of subjects who had a score on both measures being correlated.)

Table 4.7
Correlations of the LSTE-Spanish Accuracy Scores with Other Measures
(Numbers of Paired Scores in Parentheses)

	SELF-ASMT	SPANSPK	DLPTLIST	DLPTREAD	SPANLIST	ENGSPK	SPENTRAN	ENSPTRAN
MC1	.73* (65)	.76* (61)	.25 (30)	.33 (30)	.78* (25)	-.27 (18)	.14 (18)	.37 (18)
MC2	.66* (64)	.68* ;60)	.22 (30)	.29 (30)	.72* (24)	-.20 (18)	.15 (18)	.19 (18)
ACC1	.78* (64)	.80* (60)	.51* (30)	.49* (30)	.79* (24`	.01 (18)	.17 (18)	.52* (18)
ACC2	.80* (64)	.85* (60)	.44* (30)	.41* (30)	.87* (24)	-.27 (18)	-.02 (18)	.36 (18)
TOT1	.79* (64)	.81* (60)	.47* (30)	.33 (30)	.83* (24)	-.10 (18)	.17 (18)	.51* (18)
TOT2	.79* (64)	.84* (60)	.41* (30)	.29 (30)	87* (24)	-.26 (18)	.05 (18)	.31 (18)

* p < .05

We will now discuss the relationships in the table above. First, there was a significant moderate to high correlation between the LSTE-Spanish Total Accuracy score and the Accuracy score on its sections and the self-assessment, especially on the Summary section (ACC1 and ACC2). These correlations are depicted in the left hand column above. The lower, more moderate correlations with the Multiple Choice section (MC1 and MC2) support the opinion that listening ability is an important component of summary writing ability, but the fact that correlations were slightly higher with the Summary section supports the use of this task on the test. The correla ions help support the conclusion that the summary test measures abilities that the subjects involved in the study felt intuitively were needed to do summary writing.

Theoretically, as mentioned above, the ability to write a summary will depend to a large extent on one's ability to understand Spanish. The best measure of that ability available for this study is the Spanish OPI score. (It may be noted that there was a .89 correlation between speaking and listening for the 25 subjects who had both a Spanish speaking and Spanish listening ILR score). The Spanish OPI score (SPANSPK) has a high correlation with the Summary section of the LSTE-Spanish, and a moderate to high correlation with the Multiple Choice section. These correlations also help support the construct validity of the LSTE-Spanish Accuracy scores, in that such a relationship is theoretically posited between the constructs. The higher correlation with the Summary section suggests that this section places greater demands on Spanish ability than the Multiple Choice section does. This is not surprising since we saw above that the Summary section was more difficult for this sample than the Multiple Choice section (especially for Form 2), and because answering multiple choice items is more "passive" than actively writing summaries. These results also support the use of this method of testing (as opposed to using multiple choice items only) for assessing accuracy skills in summary writing ability.

The above discussion leads us to look at the relationship between the test and the ILR Spanish listening scores, available for 25 individuals. The correlations between the LSTE-Spanish Accuracy scores and the SPANLIST scores (column 5 of the table, counting from the left) are generally slightly higher than for the OPI, which again supports the construct validity of the LSTE-Spanish. On the other hand, correlations between the LSTE-Spanish and the DLPTLIST, expected to also be high, are unexpectedly low or nonsignificant. This result is not as surprising, however, when the mean, standard deviation and range of the DLPTLIST scores are examined for the 30 individuals having them. These descriptive statistics reveal that there is a ceiling effect for the DLPTLIST scores used in this sample. The DLPTLIST is designed to measure listening skills at ILR levels 1, 2, and 3. The mean of this group is 52.70 out of a total of 60 possible points and the range of the group is 39 to 60. These two figures, together with a standard deviation of 4.95, indicate that the majority of that group scored very high on the DLPT Listening test. Indeed, passing the DLPT is a prerequisite for employment in certain positions at the FBI, and most of those included in this sample had passed the DLPT listening test. Since there was very little variation and a restricted range in the DLPTLIST scores used in this study, the DLPTLIST was probably not a reliable measure for this group and hence the low correlations are not surprising. In fact, for the 27 subjects who had both a Spanish OPI and DLPTLIST, the correlation between these two was only .55. This low correlation would reflect negatively on the validity of the DLPT listening test as well if the ceiling effect of the sample scores were not kept in mind. Thus, in seeking to establish the construct validity of the LSTE-Spanish, the correlations with the DLPTLIST should be discounted.

4.4.3.2 Divergent Validity

Another criterion-related approach to establishing construct validity is to consider the correlations with measures one would expect low correlations with. This is called divergent validity. First we can look at English speaking ability, which might be related to Expression scores, but not in those scores reflective of ability to understand Spanish; i.e, the LSTE-Spanish Accuracy scores.

Although none of the correlations between the LSTE-Spanish Accuracy scores and the English OPI were significant, the fact that they were generally negative suggests that the better one's English skills, the less developed one's Spanish skills. It should be noted that the 18 individuals with English OPI scores also had Spanish OPI scores and the correlation between them was -.32 (non-significant). Of this group of 18, many were most likely bilinguals whose stronger language was Spanish. The low negative correlations between the LSTE-Spanish and the English OPI also support the construct validity of the LSTE-Spanish in a divergent manner. That is, English ability did not correlate with the LSTE-Spanish Accuracy scores, which principally measure listening ability in Spanish.

In a similar manner, we would expect low correlations between the current FBI verbatim translation measures (SPENTRAN and ENSPTRAN) and the LSTE-Spanish Accuracy scores, since the LSTE-Spanish stresses competence in listening skills and the current verbatim tests stress competence in reading. This is exactly what was found. In fact, these correlations were generally not even statistically significant.

The differences in the correlations with the LSTE-Spanish for the current Spanish into English test (SPENTRAN) and English into Spanish translation test (ENSPTRAN) also show a trend supportive of the construct validity of the LSTE-Spanish as it can be hypothesized that a verbatim written translation test is principally a measure of proficiency in the target language of the translation. Thus, those with high SPENTRAN scores can easily be weaker in Spanish than in English and should thus have lower scores on the LSTE-Spanish. Conversely, those with high ENSPTRAN scores may be stronger in Spanish than in English. ENSPTRAN should thus show a higher correlation with the LSTE-Spanish than the SPENTRAN does. Table 4.7 clearly shows this to be the case. The correlations between the Total LSTE-Spanish Accuracy scores and Spanish into English translation ability (SPENTRAN) were .17 and .05. On the other hand, the correlations between the Total Accuracy scores and English into Spanish translation ability (ENSPTRAN) were .51 and .31. This pattern of correlation provides further evidence of the convergent/divergent validity of the LSTE-Spanish Accuracy scores.

Finally, one may consider the correlations between the DLPT reading proficiency subtest (DLPTREAD) and the LSTE-Spanish. Theoretically, we would not expect these to be very high; the strength would only be in the degree to which both are measures of general proficiency in Spanish. Unlike the relationship between speaking and listening where we would expect a high correlation, we would not expect a necessarily high correlation between reading and listening. This is exactly what we see in the correlations between DLPTREAD and Total LSTE-Spanish Accuracy scores. Note, however, that these low correlations may also suffer the same problems of restriction of range as occurred with the DLPTLIST scores, although the standard deviation and the range indicate that the effect here may not be as pronounced.

It would have been useful to examine the correlations between the above measure and the LSTE-Spanish Expression scores. However, this was not possible as there were Expression scores for only 15 of the subjects for each form (as discussed in the last paragraph of section 4.3.2). Given the small number of subjects, the problem of restriction of variance since only three possible scores are awarded for Expression, and the fact that there would be even fewer subjects in paired comparisons with any other measure, correlations between Expression scores and other measures would be meaningless and thus were not calculated. Thus, it was not possible to assess the criterion-related validity of the Expression score. However, this is not important since the Expression score is a diagnostic score and not an indicator of the ability being measured by the test.

4.4.3.3 Summary

In summarizing this discussion of the construct validity of the LSTE-Spanish through the examination of convergent and divergent relationships with other measures, three things must be remembered. First, for all measures except the Spanish OPI and the self-assessment, scores were available for less than half (in many cases only one-third) of the subjects in the validation study. The correlations found above are valid only to the extent that the subsample for whom measures are available adequately reflect the entire group. We already mentioned the "ceiling effect" problem because of the unrepresentativeness of the group for which there were DLPT scores (i.e., they were all high scorers). Second, it may have been more appropriate to look at disattenuated correlation coefficients; i.e., correlations that may be expected when both measures are totally reliable. To calculate those, however, the reliabilities of all the measures must be known, and that was not the case here. On the other hand, while disattenuation tends to strengthen absolute relationships between the measures, it does not change the relative overall pattern of relationships between them. Thus, even if it were possible to disattenuate the correlations, the above discussion of the construct validity of the LSTE-Spanish would still apply. Third, the above discussion refers to the validity of the Accuracy scores only.

The evidence for the construct validity of the LSTE-Spanish as a measure of ability to understand spoken Spanish and convey that understanding in English is strong. High correlations were found with measures that would be strongly related to a high Spanish listening ability: the Spanish OPI (SPANSPK) and the Spanish Listening ILR score (SPANLIST), which are interview-based measures of speaking and listening. Low correlations were found with measures of other traits; though to the extent that they measured general Spanish ability (the English into Spanish verbatim translation test score (ENSPTRAN) and the DLPT reading subtest (DLPTREAD)), they were stronger than with those that rewarded higher English ability (the Spanish into English verbatim translation test score (SPENTRAN) and the English OPI (ENGSPK)). Problems existed with the subjects in the sample who had DLPT listening subtest scores (DLPTLIST) due to restricted range, and thus those correlations were discounted. Finally, there were strong correlations between the LSTE-Spanish Accuracy scores and the subjects' self-assessment of their summary writing ability.

Had there been LSTE-Spanish Expression scores available for a large number of subjects, we would have conducted correlational analyses similar to those carried out for the Accuracy scores. We would have expected to find very different correlations than those for the Accuracy scores discussed above. Subjects stronger in English ability would most likely have done better than those strong in Spanish ability. However, the lack of evidence of the construct validity of the Expression score is not critical, since this score is not designed to be as meaningful as the Accuracy score. The Expression score is best considered as a diagnostic subscore as treated as a "pass/no-pass" standard for decision making purposes, rather than as an attempt to accurately measure to the full extent possible the construct of English writing ability.

5. Construction of Summary Accuracy Scale Scores for the LSTE-Spanish

This section describes the rationale for the setting of the ranges of raw Ac`uracy scores to their corresponding Final Accuracy Ratings (FARs). In order to make decisions on the basis of test scores, compare test scores across forms, and interpret test scores, raw scores for Accuracy on the LSTE-Spanish must be converted to Final Accuracy Ratings.

5.1 Overview

In all of the preceding discussion of the LSTE-Spanish, raw scores have been used. However, one of the goals of the project was to be able to interpret test scores on a descriptive scale. The first step to achieving this entailed the construction of raw score to Summary Accuracy Scale (SAS) score conversion tables for the Multiple Choice section scores and the total Accuracy scores (MC + Summary Accuracy Scores) for each form of the test. These are presented in Appendix S. In this discussion, it must be kept in mind that only Accuracy scores are involved. The Expression scores (Deficient, Functional, and Competent) are not converted and are always reported separately from Accuracy scores.

Construction of the scale score conversion table is an attempt to give interpretative meaning to the LSTE-Spanish raw scores for Accuracy. In addition, it enables the comparison of scores across forms. Conversion into scale scores takes into account differences in test length and test difficulty. Thus, a comparison of results for Accuracy across test forms and subtests must only be made in terms of the scale scores, and ultimately on the basis of the Final Accuracy Ratings.

5.2 The Selection of the Criterion Variable

Since one of the goals of the project was to provide translation ability scores based on a descriptive scale, it was necessary to select an existing ILR score that would help anchor

LSTE-Spanish scores in the ILR scale. This was found during the validation study (see section 4) in the Spanish OPI score (SPANSPK), which correlated quite highly with the LSTE-Spanish Accuracy scores and was available for 60 of the 66 subjects in the validation study. In addition to the OPI score, we had available for most of the subjects a measure that also correlated highly with the LSTE-Spanish Accuracy score, namely, the subjects' self-assessment scores (SELF-ASMT). It should be noted that these two measures also correlated well with each other (.84), showing that they were measuring similar constructs. Plots of the LSTE-Spanish Accuracy scores against both of these variables showed that the fit between the self-assessment score and the test scores in the critical ranges of ILR 1 to 4 was actually better than that for the OPI. That is, a small but significant group of subjects in this range performed much higher on the LSTE-Spanish than their OPI scores alone would predict. This may have been due to the fact that for these subjects the OPI score was outdated, since OPI scores were not current with taking the test. Some of them were over three years old. The self-assessment data, however, was current with the test taking; subjects completed the self-assessment directly before taking the tests.

In light of the above, it was decided to use a composite of the OPI and the self-assessment as the best indicator of current summary writing ability. Doing so had the beneficial effect of both adjusting out-of-date OPI scores to more accurately represent abilities at the time of test taking and adjusting OPI speaking scores to better reflect summary writing ability.

To form a composite criterion score for each subject, first all examinees who were missing any LSTE-Spanish, OPI or self-assessment scores were eliminated from the data set. This left 58 of the 61 subjects with Spanish OPI scores. Second, to ensure equal weighting in the composite score, the OPI and the self-assessment scores were transformed into standardized t scores;

i.e., they were linearly transformed to have a mean of 50 and a standard deviation of 10. Once this was done, the third step was to add the two standardized scores together. Finally, this total score composite was scaled through a linear transformation to correspond back to the ILR scale. This transformation used two anchor points. The first was the highest possible raw score on the two measures (5.0 on the OPI and 12 on the self-assessment, which was equal to 123.76 on the total score composite). This was assigned to a 5.0 on the Summary Accuracy scale. The second anchor was the "minimally competent" score (2+ on the OPI and 6, i.e. Functional on the self-assessment, equal to 68.64 on the composite). This was assigned to a 2+ (2.8) on the Summary Accuracy scale.

The formula for a linear transformation is

scale score = A x raw score + B

where A is the slope (i.e., scaled score2- scaled score1/raw score2 - raw score1) and B is the intercept (i.e., scaled score2 - A x raw score2). By substituting the equivalencies given above, the following equation was derived for converting the composite scores to the scale score:

scale score = (.0399 x composite score) + .06038

In this way each examinee received a Summary Accuracy Scale score for accuracy in summary writing ability. Appendix Q shows each examinee's OPI, self-assessment (SA), and SAS score. An examination of the scores shows how the transformation brought the scores of those with relatively lower OPI scores but higher self-assessment scores up on the SAS and those with higher OPI scores but relatively lower self-assessment scores down. For example, compare examinee 45, with an OPI of 3.8 and a self assessment of 10 and who received an SAS score of 4.10, with examinee 15, with an OPI of 4.8 and a self-assessment of 8, who received an SAS score of 3.99, indicating a very similar summary writing ability despite differences in the OPI performance.

The logic of the transformation from straight OPI scores to

SAS scores may be explained further as follows. It was expected that of those scoring at the top of the ILR range (4+/5), almost all were native speakers of Spanish whose skills in English may have made them more hesitant to give themselves a high rating on the self-assessment. In fact, of the 20 subjects in the validation study sample scoring 4+/5 on the OPI, only two who were rated at 4+ were higher on the SAS score and only two who were rated at 5 on the OPI were still at 5 on the SAS score. The other 16 had SAS scores lower than their OPI scores. The SAS scores for this group of 20 high OPI scorers ranged from 3.98 to 5. It is reasonable to assume that subjects at the other end of the OPI scale (0+/1 range) were native English speakers. Since at this level passive skills in listening comprehension may exceed active skills in speaking, it would be assumed that their scores on the SAS may be slightly higher than their OPI scores. Indeed, as a result of the linear scale transformation, the lowest subjects in the sample (those who scored 0.8 on the OPI and 3 on the self-assessment) received a 1.38 on the SAS. This score is not unreasonable. Eleven of the 17 subjects in the (0+/1) category scored between 1.38 and 1.45 on the SAS; 16 (94%) of them had SAS scores at 2 or below, with the final subject (examinee 78, who got a 6 (Functional) on the self-assessment scoring at about 2.12).

To see whether this score fit the test data better, we can compare the relationship between the individual parts. This is presented in Table 4.8 below.

Table 4.8
Correlations of Spanish OPI, Self-Assessment (SA) and Composite OPI+SA Score Converted to the Summary Accuracy Scale (SAS) with the LSTE-Spanish Accuracy Raw Scores

	OPI	SA	SAS
MC1	.75	.75	.78
MC2	.68	.71	.72
ACC1	.79	.81	.84
ACC2	.85	.83	.87
TOT1	.81	.82	.85
TOT2	.83	.83	.87

Note: n = 58 (all examinees with complete data)

Table 4.8 shows that using the composite score for SAS gave a slightly better fit to the test data and thus provides an adequate foundation for building a score conversion table.

Another way to ensure the appropriateness of the scale score is to examine it on the basis of the score required on the self-assessment in order to bring the SAS score to a 2.8 or above level. With an OPI score of 1, one would need to have had a self-assessment score of 9 (= Competent) to get a score of 2.8 on the SAS. With an OPI score of 1.8 or 2, one would need a score of 8 (1 Functional and 2 Competents) to get a score of 2.8 on the SAS. It is unlikely that this would happen if both measurements were taken concurrently. Thus, the SAS score conversion has intuitive logic and meaning that is useful for decision making.

We can now see if the second goal in using a composite score was met; i.e., that of updating out-of-date OPI scores. For our purposes, we are most concerned with those subjects in the 2 to 3 range. There were 9 in the validation sample. A preliminary regression analysis revealed that for two of these subjects, their predicted scores on the LSTE-Spanish were much lower than the score they actually achieved. Both were at level 2. When their self-assessment was analyzed, one scored 8 (two Competents and one Functional) and the other 10 (two Competents and one Superior). In this case, their actual current ability was probably much greater than an OPI of 2.0 would suggest. Their SAS scores were 2.99 and 3.47 respectively. If only the OPI were used as a criterion variable, then these subjects would have negatively influenced the data to a greater extent than by using their SAS scores.

5.3 Outliers Detected and Removed

The preliminary examination of the raw score data revealed that there were some highly inconsistent cases. This is one reason why the SAS scores were developed rather using than a straight OPI score. However, it remained to be seen whether there were still any outliers in the set whose test performance behavior can not be explained by using the SAS score. Inclusion in the data set to convert LSTE-Spanish scores into SAS scores might jeopardize the usefulness of the results for score interpretation and decision making. To detect any outliers, a regression to predict SAS from each of the LSTE-Spanish Accuracy subtest scores and Total Accuracy scores was run. Those cases which had the largest residuals were marked. This occurred sporadically across 13 of the 58 subjects. However, for two subjects there were consistent problems. After further analyses, it was clear that one subject's test performance was surprisingly and consistently higher on one form of the test than the other. In the other case, the test scores were remarkably high for the reported OPI score and the self-assessment score. These two subjects were thus deleted from the data set before proceeding to scale the test.

5.4 Development of Raw Score to Scaled Score Conversion Tables

To develop a conversion table of raw LSTE-Spanish scores to SAS scores, a regression was run to predict SAS scores from

Multiple Choice section and total Accuracy scores. From these regressions, four regression equations were made. These equations were then used to predict SAS scores from Multiple Choice section scores and from total Accuracy scores. These four conversion tables are presented in Appendix S. The following comments must be noted:

1. For the Multiple Choice sections, a score of 10 or below can be achieved by chance. Thus, there is no SAS equivalent for those scores.

2. These conversion tables take into account differences in the number of items on the test forms and differences in difficulties. In other words, the scale score represents score equivalencies on the two forms. This can be seen by looking at the mean scores on each test (n = 56):

Test	Rounded Mean Raw Score	SAS
MC1	29.00	3.33
MC2	27.50	3.33
TOT1	72.50	3.32
TOT2	64.00	3.32

3. Although it would be possible to convert from both the Multiple Choice section and the total Accuracy score to the SAS, the most accurate measurement is of course on the basis of the total score, for two main reasons. First, the total score, which is a composite of the two section scores, contains more variance and a wider spread of scores than the Multiple Choice section score alone. Second, the total score correlated more highly or as highly with SAS than did the Multiple Choice section.

4. As is true whenever regression equations are used, the most accurate conversions will be around the mean of the scales (see chart above). This is close enough to the cut-off score of 2.5 to be confident of its usefulness at that range. It will be less accurate at the extreme ends of the range.

5. The Multiple Choice section is best used as for screening purposes (see below).

5.5 Defining the Final Accuracy Rating Boundaries

The Final Accuracy Rating scale is based on the six descriptions of summary writing ability: No Ability, Severely Deficient, Deficient, Functional, Competent and Superior. (It may be noted that the self assessment categories used in the above analysis were approximately equivalent to the categories labeled Deficient to Superior.) The scale was developed based on two sources of input. First, we considered the discussion of accuracy (misinterpretations, omissions, and additions) in the FBI/CAL translation skill level descriptions. Next we considered the range of performance of examinees in terms of the summaries they wrote. The six descriptions on the Interpretion of the Final Accuracy Rating thus represent holistic performance descriptions that were written in reference to the translation skill level descriptions and to natural performance groupings within the sample tested. While the correlations between the Accuracy and the Multiple Choice sections were high (.83 and .79 on Forms 1 and 2 respectively), only the summaries represent performance samples from which a performance description can be extracted. Still, given the high correlation between the two sections, and the similarity of the listening stimuli and the type of information tested by the multiple choice (main topic, key points, and supprting details), it is appropriate to use the performance description to interpret performance on the total test; i.e., the multiple choice section and the total Accuracy score combined.

The cutting point for the rating of No Ability was developed considering the chance score on the Multiple Choice section and the number of points that could be earned by examinees who only identified the name of the speakers in the conversations in the Summary section. There were forty multiple-choice items; thus the chance score on this section is 10. This level of performance or lower represents no ability. There were two speakers in each of the three conversations for a total of six points that could have been obtained by an examinee at the No Ability level on the summary writing section. Thus, the cutting score for the top of the No Ability range is a raw score of 16 points on both forms. The cutting score for the remaining descriptions is the point at which the corresponding Summary Accuracy Scale score exceeds .50. Thus, the remaining cutting scores are the raw scores that are equivalent to an SAS score of 1.50, 2.50, 3.50 and 4.50. The range for the Incompetent category goes from the cutting score for No Ability to 1.49; for the Deficiency category it goes from 1.50 to 2.49; for the Functional category from 2.50 to 3.49; for the Competent category from 3.50 to 4.49; and finally for the Superior category it goes from 4.50 to 5.0.

It may be useful at some point in the future for the FBI to perform a cross-validation analysis of the Final Accuracy Rating descriptions (Appendix R). In other words, an analysis could be carried out of the performance of examinees in each category on the FAR scale in terms of a) whether or not they successfully communicated the topics of the conversations, b) the average number of key points, and c) the average number of supporting points they identified in their summaries. These actual mean performance levels could then be compared with the FAR scale descriptions.

5.6 Using the Multiple Choice Section as a "Screen"

The Multiple Choice section of the LSTE-Spanish may be used to screen out individuals for whom the Summary section of the test is inappropriate; that is, examinees would not be likely to have a total Accuracy score at a 2.5 or above on the summary accuracy scale (Functional or above on the FAR). In this case, the most serious error to make in using the Multiple Choice section score is to make a decision to exclude someone from taking the Summary section rather than to give the Summary section to someone who may not ultimately receive a FAR of Functional. To determine the cut-off score on the Multiple Choice section, we need to first determine the raw score on the Multiple Choice section that corresponds to a Functional score (2.5 on the SAS). Once this is found, we then need to determine the lowest possible raw score one could get on the Multiple Choice section while, given measurement error, still having a statistical possibility of scoring at that cut-off score level.

The raw score on the Multiple Choice section of Form 1 that most closely corresponds to a passing score of 2.5 is 24 (2.59); on Form 2 it is 21 (2.50). Given the reliability of the two tests at .86 and .88 respectively and the variances of the validation study sample (see section 4.3.1), the standard error of measurement (SEM) for Form 1 is 2.59 and for Form 2 it is 2.68. Thus, the 95% confidence interval around the passing score would then be:

MC1 24 - 2 x 2.59 to 24 + 2 x 2.59 = 18.82 to 29.18
MC2 21 - 2 x 2.68 to 21 + 2 x 2.68 = 15.64 to 26.36

This means that an examinee scoring 19 or below on Form 1 of the Multiple Choice section or 16 or below on Form 2 has less than a 2.5% probability of having a "true" raw score of 24 or 21, respectively, on each form, which correspond to a 2.5 on the SAS. These, then, would be the cut-off scores. Examinees who score below this level on the Multiple Choice section of the LSTE-Spanish either need not take the Summary section, or if they already have, that section need not be scored.

Using these cut-off scores would still leave in many examinees who may not ultimately achieve a Finaly Accuracy Rating above Functional; however, the chance of excluding a candidate who might achieve a Functional is slim.

As a final comment, it is obvious that scores on the Multiple Choice section cannot predict Expression scores. That is, a candidate may achieve a passing score on the Multiple Choice section (and on the Final Accuracy Rating), yet ultimately not pass the LSTE-Spanish on the basis of a Deficient Expression score. The Multiple Choice section of the LSTE-Spanish is not intended as to screen out such candidates.

APPENDIX A

MULTIPLE CHOICE SECTION TEST BOOKLET

(SELECTED PAGES)

NAME ______________________________
Last   First

DATE __________

SPANISH LISTENING SUMMARY TRANSLATION EXAM

PART A: MULTIPLE CHOICE

FORM 1

This test is for official use only; do not divulge any information contained herein.
Do not duplicate any portion of this test. Do not show to unauthorized persons.

FIELD OFFICE ______________________________

TEST NO. _____

PART A: MULTIPLE CHOICE ITEMS

In this exam you will hear a series of conversations. Some of the conversations are related to each other. In Part A you will hear each conversation only once. You will find several questions in your test booklet based on each conversation. Each question is followed by four possible answers. Before you hear the conversation, you will be given an opportunity to briefly sc n the questions and possible answers. This will help you know what type of information to listen for.

As you are listening to the conversation, you may scan the questions again and mark your choices in the test booklet. Do not be distracted by slang or phrases which are unfamiliar to you. Instead, concentrate on extracting only the information needed to answer the questions. You do not need to understand every word to answer the questions correctly; however, you will need to be alert and attentive.

After listening to the conversation, you will have a brief period of time to review your choices and transfer your answers to the answer sheet. Make sure you have selected the best answer for each question, based on what you heard in the conversation. On your answer sheet, find the number of each question and fill in the space that corresponds to the letter of the answer you have chosen.

EXAMPLE

1. What day will Teresa and Raúl meet?

   A. Monday
   B. Thursday
   C. Saturday
   D. Sunday

2. At what time will they meet?

   A. 5:00
   B. 6:00
   C. 7:00
   D. 8:00

3. Why are they meeting?

   A. To have dinner
   B. To discuss business
   C. To go dancing
   D. To close a deal

DO NOT TURN THE PAGE UNTIL YOU ARE ASKED TO DO SO.

APPENDIX B

SUMMARY SECTION TEST BOOKLET

(SELECTED PAGES)

NAME ______________________________
Last First

DATE __________

SPANISH LISTENING SUMMARY TRANSLATION EXAM

PART B: SUMMARY

FORM 1

This test is for official use only; do not divulge any information contained herein. Do not duplicate any portion of this test. Do not show to unauthorized persons.

FIELD OFFICE ______________________________

TEST NO. _____

PART B: SUMMARY ITEMS

In this part of the test, you will hear an example conversation followed by three additional conversations. These conversations will be similar to those you listened to in Part A. This time, however, you will hear each conversation twice, and you may take notes on the content of the conversations. After each conversation, you will be asked to write a summary in ENGLISH of the important information.

You will not use an answer sheet during this part of the listening exam. Instead, there is space provided in your test booklet for you to jot down information given by the speakers. As you listen to a conversation for the first time, write down important information in the space marked "NOTES" corresponding to that conversation. Then, as you listen a second time, make any corrections to your notes that may be necessary, or add details you may have missed.

Important information relates to the general purpose of the conversation and supporting details including names, dates, times, places, or amounts. The conversations vary in the amount of concrete information they contain. If a conversation deals with more abstract topics, make sure you identify the general topic and primary supporting points.

After each conversation, you will have a limited amount of time to write a summary of the conversation in as much detail as possible. The conversations increase in length from approximately one to three minutes. The amount of time you are given to write a summary will depend on the length of the conversation. Before you begin writing, you will be informed of how much time you will be given to complete each summary. You will also be advised when there is one minute remaining to complete your summary. Write the summary in the space marked "SUMMARY" corresponding to a particular conversation. Be sure to write legibly in complete sentences. (Do not worry about the legibility of your notes as only the summary will be scored.) Include only information you have gleaned from the conversation; do not add any of your own assumptions or inferences.

EXAMPLE

NOTES

SUMMARY

EXAMPLE

SUMMARY

Meche calls Paco to confirm that she's arriving Sunday morning at 7:30 a.m. on United flight #517. She asks him to wait for her in the baggage area because she's bringing a lot of luggage. He says he will be there, but in case he's not, to ask someone to help with her bags and to wait for him outside. She agrees and says she'll either see him in the baggage area or outside with the "goods."

EXPLANATION

Notice the important information that was included in the summary:

The names of both parties in the conversation have been noted. Specific information about Meche's arrival, including the day, time, airline, and flight number has been recorded. In addition, the discussion between Meche and Paco about her luggage has been summarized. Finally, the location where Meche and Paco agree to meet has been identified. Note that the sum.mary also includes Meche's reference to the "goods."

Notice that the summary has been written in full sentences and in paragraph form, rather than as a list. Remember that your score will depenc not only on the content of your summary, but on the way it is written as well.

APPENDIX C

EXAMPLE SUMMARY CHECKLIST

SPANISH LISTENING SUMMARY TRANSLATION EXAM

EXAMPLE SUMMARY CHECKLIST

_____ 1. **Meche**
_____ 2. **Paco**
_____ 3. She requests that he **pick her up** at the airport
_____ 4. on **Sunday**
_____ 5. **morning**
_____ 6. The flight is **United**
_____ 7. **#517**
_____ 8. arriving at **7:30**
_____ 9. He **agrees** to come,
_____ 10. but **if he's late**
_____ 11. she should ask someone to **help with the bags**
_____ 12. and **wait** for him **outside.**
_____ 13. She would prefer **he meet her** in the baggage area
_____ 14. because she has **a lot of things.**
_____ 15. She'll **feel more at ease** if he's there.
_____ 16. She will either meet him in **the baggage area**
_____ 17. or **outside**
_____ 18. with the goods/**merchandise.**

_____ _____ (Speakers set up **meeting at airport.**)
Total Topic
(18) (2)

Accuracy _____ Expression _____

APPENDIX D

SUMMARY EXPRESSIOI! GUIDELINES

SPANISH LISTENING SUMMARY TRANSLATION EXAM

SUMMARY EXPRESSION GUIDELINES

DEFICIENT — At this level, the summary may rely on basic grammatical and syntactic structures and exhibit little attempt to connect sentences. It may contain several errors in basic grammar structures, spelling, or punctuation as well as Spanish words or false cognates.

FUNCTIONAL — At this level, the summary may contain run-on sentences, sentence fragments, or awkward and ambiguous wording (including inappropriate register) that interferes with the presentation of ideas. It may contain errors in complex grammar structures and common errors in spelling or punctuation.

COMPETENT — At this level, the summary conveys the information in a logically organized manner (which may differ from the original order of presentation), but the phrasing may be choppy or wordy. While the use of grammar is generally accurate, the summary may exhibit a telegraphic note-taking style, including omission of articles and subject pronouns. Spelling and punctuation are generally accurate and vocabulary is adequate, although not extensive and precise.

APPENDIX E

FINAL ACCURACY RATING
SCORE CONVERSION TABLE

LSTE-SPANISH FINAL ACCURACY RATING

CONVERSION TABLE[6]

Form 1 Raw Total	Form 2 Raw Total	Final Accuracy Rating
0 - 16	0 - 16	No Ability
17 - 32	17 - 25	Severely Deficient
33 - 54	26 - 46	Deficient
55 - 76	47 - 67	Functional
77 - 98	68 - 89	Competent
99 - 114	90 - 111	Superior

[6] The difference in conversion scores between the two forms of the LSTE reflects the difference in total Accuracy points possible on each form and the fact that Form 2 is more difficult than Form 1.

APPENDIX F

CHARACTERISTICS OF MONITORED CONVERSATIONS

CHARACTERISTICS OF MONITORED CONVERSATIONS

[Quality of recording often isn't good--voices sometimes difficult to hear, background noise]

1. Frequently call collect, use first names only

2. Sociolinguistic characteristics:
   a. A lot of man to girlfriend talk
   b. Politeness conventions are usually respected, including preliminary small talk, unless a delivery is being confirmed (in which case conversation is very brief)
   c. TexMex go through elaborate verbal greeting and leave- taking rituals
   d. Tone varies from nervous and hurried to relaxed and friendly, humorous
   e. A lot of obscene language (Cuban: encabronarse; Mexican: chingar)
   f. Interlocutors may be from different countries resulting in a mixture of diaiects

4. Forms of address or to refer to 3rd person:
   a. Use of nicknames, i.e. mijo, hermano, mano, negro, gordo, flaco, doctor, ingeniero, el inge, el licenciado (for higher-up), viejo; Mexican: compadre, compa/cumpa, mero-mero; Puerto Rican: jíbaro; Chicano: mi sangre, el bro
   b. Men use diminutives with women, i.e. negrita

5. Speak in very general terms: it, something, that, that thing, that office, the other ones, there, here, little brothers, grandmother, nephew, compadre, pariente, cifras, units, parts, points, details, products, European ones, the car, friends, aquel, aquella, paquete

6. Code, argot:
   a. Drugs: rubios, el polvo, la harina, botas, perlas
   b. Money: los verdes, la lana
   c. Police, government agents: los Fedes (Mexican agents); gueros

7. Colloquial expressions: lo va estirar; se metió en un berenjenal; lo van a mandar a ver a San Pedro

8. Frequent topics: times, days, dates (arranging for meetings, arrivals, departures. phone calls); verifying phone numbers; checking amounts and prices; making travel plans, methods of payment (cash, in kind, check, transfer); whether or not someone is trustworthy; el biper [beeper]; amounts (kilos, cuartos, octavos)

9. Confirmation devices: chévere, ok, fenómeno, ándale pues (ándale also used during conversation to indicate comprehension)

APPENDIX G

LSTE-SPANISH EXAM FORMAT

(PILOT VERSION)

PILOT VERSION

SPANISH LISTENING SUMMARY TRANSLATION EXAM FORMAT

Multiple Choice Items

TOPIC	FORM A	TIME	FORM B	TIME
Example	26[1] mf[2]	30s	26 mf	30s
General	27 mm	30s	28 mm	28s
Terrorism	31 mm	18s	32 mm	18s
Drugs	1a mf	1m 38s	2a mm	1m
	1b mf	1m 28s	2b mm	1m 5s
	4 mm	42s	2c mf	51s
	6a mf	1m 53s	3 mf	49s
	6b mf	1m 14s	13 mf	2m
	8 mm	1m 50s		
Fraud	11 mf	1m 30s	9 mf	2m 33s
			14a mf	1m 53s
			14b mm	2m 30s
Immigration	12 mm	1m	15 mm	1m 18s
	16 mf	1m 45s		

Multiple Choice Section Subtotals

# of conversations (including one example)	12	12
Time	14m 20s	15m 15s
Interlocutors	5 mm 7 mf	6 mm 6 mf
# of multiple choice items developed[3]	60	54

[1] Each scenario was assigned a number when originally drafted.

[2] Sex of interlocutors: mm = two males; mf = a male and a female

[3] More items have been developed than will appear in the final version of the exam, as it is anticipated that several items will be deleted after piloting.

Summary Items

TOPIC	FORM A	TIME	FORM B	TIME
Example	25 mf	1m 10s	25 mf	1m 10s
General	30 mm	50s	29 mf	47s
Terrorism	17 mm	2m 5s	18 mm	2m 3s
FCI	23 mm	4m 50s	24 mm	3m 40s

Summary Section Subtotals

# of conversations (including one example)	4	4
Time	8m 57s	7m 40s
Interlocutors	3 mm 1 mf	2 mm 2 mf

Exam Totals

# of conversations	16	16
Time	23m 17s	22m 55s
Interlocutors	8 mm 8 mf	8 mm 8 mf

APPENDIX H

SUMMARY SCOPING GUIDE

(PILOT VERSION)

PILOT VERSION

SPANISH LISTENING SUMMARY TRANSLATION EXAM

SUMMARY SCORING GUIDE

SUBSTANTIVE ACCURACY

No ability	No response or fails to represent overall topic accurately. Provides no substantive information beyond names of speakers.		0	5
Incompetent	Fails to represent topic accurately. Contains frequent misinterpretations, omissions, and/or misleading additions; about a fourth of the key items are correctly reported.		10	15
Deficient	May not represent topic accurately. Contains many misinterpretations, omissions, and/or misleading additions; about half of the key items are correctly reported.		20	25
Functional	Represents topic accurately; however, contains misinterpretation, omission and/or misleading addition of several key items. May contain a number of supporting details.	30	35	40
Competent	Reports all or almost all key items accurately and many supporting items as well; no misleading additions.	45	50	55
Superior	No misinterpretations or omissions of key items, although some nuances may or not be conveyed. Might include information regarding the mood and/or relationship of the speakers.	60	65	70

GRAMMAR

No ability	No response.		0
Incompetent	Majority of grammar structures are incorrect.	1	2
Deficient	Contains many errors in basic grammar structures.	3	4
Functional	Relies on basic structures. Some errors in complex grammar structures, but few in basic structures. May contain sentence fragments.	5	6
Competent	Uses generally accurate grammar. (Telegraphic note-taking style, including omission of articles and subject pronouns, is acceptable at this level.)	7	8
Superior	Uses complex structures, almost always without error.	9	10

SPELLING/PUNCTUATION

No ability	No response.	0
Incompetent	Many errors in spelling and punctuation.	1
Deficient	Several errors in spelling and/or punctuation.	2
Functional	A few common errors in spelling and/or punctuation. May contain run-on sentences.	3
Competent	Generally accurate spelling and punctuation.	4
Superior	No errors in spelling or punctuation.	5

VOCABULARY

No ability	No response or response entirely in Spanish.		0
Incompetent	Contains many Spanish vocabulary items.	1	2
Deficient	Contains Spanish vocabulary items, or false cognates translated literally from Spanish.	3	4
Functional	Contains awkward or ambiguous wording (including inappropriate register).	5	6
Competent	Vocabulary is adequate, though not extensive and precise.	7	8
Superior	Extensive and precise vocabulary.	9	10

ORGANIZATION

No ability	No response.	0
Incompetent	Consists of isolated words or phrases.	1
Deficient	Evidences little attempt to connect sentences or fragments smoothly.	2
Functional	Sentence fragments, run-on sentences, or awkward phrasing may interfere with organization of ideas.	3
Competent	Conveys information in a logically organized manner (which may differ from original order of presentation). However, phrasing may be choppy or wordy.	4
Superior	Conveys information concisely in a logical manner. Uses a variety of sentence patterns and organizational devices (such as transition words and phrases).	5

APPENDIX I

LSTE-SPANISH EXAM FEEDBACK QUESTIONNAIRE

(PILOT STUDY)

PILOT STUDY QUESTIONNAIRE

SPANISH LISTENING SUMMARY TRANSLATION EXAM

We would very much appreciate your answers to the following brief questions concerning the listening exams you have just taken. Your comments will help us to identify aspects of the exams which need to be improved. Thank you for your cooperation.

PART A (MULTIPLE CHOICE ITEMS)

1. Were the directions for Part A clear?

( ) Yes ( ) No

COMMENTS:

2. Were the example items helpful?

( ) Yes ( ) No

COMMENTS:

3. Do you have any comments about any of the conversations in Part A?

4. Were the pauses between questions about the right length?

( ) too short
( ) about right
( ) too long

COMMENTS:

PART B (SUMMARY ITEMS)

1. Were the directions for Part B clear?

   ( ) Yes ( ) No

   COMMENTS:

2. Was the example summary helpful?

   ( ) Yes ( ) No

   COMMENTS:

3. Do you have any comments about any of the conversations in Part B?

4. Were the pauses for you to write the summaries about the right length?

   ( ) too short
   ( ) about right
   ( ) too long

   COMMENTS:

5. Please use the space below to comment on any aspects of the exams that were not covered in the preceding questions. We would appreciate any suggestions as to how these exams might be improved.

Thank you again for your help.

APPENDIX J

LSTE-SPANISH EXAM FEEDBACK QUESTIONNAIRE RESULTS

(PILOT STUDY)

PILOT STUDY QUESTIONNAIRE

SPANISH LISTENING SUMMARY TRANSLATION EXAM

RESULTS

PART A (MULTIPLE CHOICE ITEMS)

1. Were the directions for Part A clear?

   (98%) Yes (2%) No

2. Were the example items helpful?

   (95%) Yes (5%) No

3. ...

4. Were the pauses between questions about the right length?

   (15%) too short
   (51%) about right
   (34%) too long

PART B (SUMMARY ITEMS)

1. Were the directions for Part B clear?

   (100%) Yes ( ) No

2. Was the example summary helpful?

   (67%) Yes (33%) No[1]

3. ...

4. Were the pauses for you to write the summaries about the right length?

   (18%) too short
   (76%) about right
   (6%) too long

5. ...

Thank you for your cooperation.

[1] This reflects the fact that about 20 (43%) of the examinees (GM) did not have an example summary in their test booklets, due to an error in duplicating the materials.

GM: Also not enough writing space.

OTHER: In conversation 3, there wasn't enough time to finish the summary.

There isn't enough space to write the last summary. I ran out of time on all but one. I noted when my time ran out, then continued my answers.

5. **Please use the space below to comment on any aspects of the e:am that were not covered in the preceding questions. We would appreciate any suggestions as to how these exams might be improved.**

FBI: To allow to mark the answer in the test booklet.

FBI: It is much better than the one I took to get the job. Good job.

UM: Think examinees should be encouraged to speak <u>and</u> write <u>in</u> <u>Spanish</u>.

UM: Because hispanics tended to speak quickly people being given test should be allowed to mark answers on question sheet.

UM: It was fun.

UM: Wrote in pen on 2nd part.

CAL: You're right. These are clearer than anyone will <u>ever</u> hear. But for a test, that's good, my hand and arm were <u>wiped</u> <u>out</u> at the end of this -- even at the end of Form 1.

CAL: In the summary part, it would be nice to have the time remaining announced periodically: "There are 5 minutes left," "There are 2 minutes left," "There is 1 minute left."

It would also be good to have pp. 18 and 19 facing each other. I tore out my p. 19 because I didn't want to be flipping back and forth from my notes all the time.

No pause after second beep. Insert several seconds after each paragraph and each beep.

It would help to <u>do</u> a sample summary before the real ones, since one learns strategies after the first item. I feel I could have done better on the first one. Also, these items obviously depend on note-taking ability a lot. This is distinct from comprehension ability. Do you want to be testing note-taking ability?

CAL: 1) Background noise throughout on both tapes.

2) One summary section conversation seems to have a problem at the beginning. No second name is evidenced.

GM: They seemed kind of difficult but if you are testing for fluent speaking and listening the test is good.

GM: Be clear whether you want people to guess.

GM: Accurate.

GM: They were fine.

GM: Slow down the speaking a little bit.

GM: Always do in 2 parts -- very tedious and your mind drifts after only a short time -- especially Part A.

GM: I wish I had understood more of the conversations -- it would have been more interesting.

GM: A break between sections ought to be built into the administration of the test.

OTHER: The explanations are boring. This part of the test ("fill in") should take less time.

PILOT STUDY QUESTIONNAIRE

SPANISH LISTENING SUMMARY TRANSLATION

COMMENTS

General Comments:

CAL[1]: It was fun!

PART A (Multiple Choice)

1. **Were the directions for Part A clear?**

FBI[2]: There's a lot of time in between the directions.

FBI: Make clear that one cannot even write a check mark in the booklet.

FBI: No problem.

UM[3]: Initial instructions were too slow for the native English speakers. I suggest that there be two versions of the tape, one for native and one for non-native speakers.

UM: It did not mention to fill in sex or grade level.

UM: Wordy

UM: There was no need for the speaker to say "Question 10" as we answered. It was distracting. We were all way ahead trying to answer the questions while the info was fresh.

UM: It was wordy in parts, such as when he had to say every letter on the answer sheet "A, B, C, D, E" etc.

CAL: It seems that the directions are quite prolonged.

CAL: There should be a pause between first and second paragraphs. Now sounds like these have been spliced together. No pause is uncomfortable. Also, insert pauses between paragraphs on p. 2.

[1] Identifies a CAL staff member who participated in the trialing of the LSTE-Spanish.

[2] Identifies an FBI staff member.

[3] Identifies an intermediate Spanish language student at the University of Maryland.

CAL: Maybe it would be good to say whether it's better to guess or leave spaces blank.

CAL: I insert more of a pause before and after the beeps so the supervisor has time to get to the machine. (Also between paragraphs, in example instruct., expl.)

CAL: It seems that the directions are quite prolonged.

GM[4]: Very clear.

GM: You should state whether you want people to guess or just leave it blank if they don't have a good idea what they heard. It's also difficult (#2 & #7) to tell who has called who.

GM: The speakers talked a little bit too fast.

GM: It was only after I strained my brain trying to understand what you meant. Awkward phrasing.

GM: Repeat the conversations once each, because if you are listening to a phone call, you are also taping it and can hear it again.

OTHER: A greater variety of accents would help determine true ability to understand

It is supposed that those who take the exam do speak English, so the directions should take less time. The directions are too repetitive. You could test the students' memory by their ability to remember them.

Too slow

2. **Were the example items helpful?**

FBI: Good examples.

UM: Too much time between examples.

UM: False expectations of a slower conversation.

CAL: Provide space for actual answering of examples, or indicate that a written response is not necessary.

CAL: Very clear. Good example and good explanation.

CAL: Provide space for actual answering of examples, or indicate that a written response is not necessary.

[4] Identifies a beginning Spanish language student at George Mason University.

CAL: Very much.

GM: They were helpful.

GM: The directions were clear enough.

GM: For part B there was no example summary.

GM: But I had difficulty following the conversations.

GM: They helped show you what specific items the estions would ask.

GM: You should have allowed us a space to actually do and write the example through.

OTHER: Helpful but not necessary

Depende de la persona o ser examinado. Absolutamente no para el entrenamiento de personal diplomático.

3. **Do you have any comments about any of the conversations in Part A?**

FBI: I had a difficult time understanding their slang in terms of money and suspect it may be due to my own inexperience.

FBI: I find the conversations to have been substantively correct and natural pertinent, but they were spoken too fast for my level of comprehension.

FBI: One can easily tell what we do: wiretaps. I don't think people other than FBI personnel should know this.

UM: Tape appeared garbled in parts

UM: Why did they use Spanish slang?

UM: There were several questions to answer and for individuals who are not native speaks it could be difficult because the hispanics were speaking very fast

UM: Some of the test takers were laughing loudly during the conversation. THis made it hard to focus and understand.

UM: The people spoke a little too fast and some of the words were hard to comprehend.

CAL: Some are much more difficult than others. I found the first ones more difficult, because they were so short and fast I missed them.

CAL: Sounds like authentic street language. They sound extremely authentic. Great job! They make me want to make me improve my Spanish!

One thing though -- it seemed uke there were several "story lines" running through the options. That is, once you answer the first question, you are drawn to the options in the next questions that are congruent with your first choice. The items don't seem independent.

CAL: The first 2 were quite unclear and a number had a great deal of regional slang. Do applicants need to know the slang to be accepted? Also, the slang all seemed to be from the same region, "andale," "ciaou," "chingadas," etc.

GM: I have had three semesters of Spanish and they still seemed rather difficult.

GM: The male-female ones are easier to decipher since it's obvious who called who. The male to male aren't as easy.

GM: Incidentally, I've received A's in Spanish. Do people really speak that fast? No way for an Intermediate level Spanish student.

GM: They went kinda quick and some were rather mumbled (but that's the way we all talk, right?)

GM: Just that they talked too fast and used words we didn't know.

GM: The voices were not always very clear.

GM: They were much faster and different from the things we've learned in a classroom.

GM: I am not used to the speed of the conversation or the various inflections. Some words, possible slang, were completely unfamiliar !

GM: They spoke very fast and were somewhat difficult to comprehend.

OTHER: Some accents were hard to understand.

Absolutamente coloquiales--No práctica para medir un nivel de diálogo social.

Most of the conversations dealt with drugs and its terminology. I found this slang difficult to follow.

Form 1, Q's 34-38: I found this conversation very difficult to follow.
Form 2, Q's 23-26: I found this conversation very difficult to follow.

4. **Were the pauses between questions about the right length?**

FBI: My only concern is that the time frame creates a problem that is quite independent from the language skill of the candidate. Mistakes may be due to the difficulty in remember details, not in understanding the conversation.

FBI: Are you testing recall or language understanding? Why not let me circle the correct answer on the booklet as I hear the conversation. This will reflect a true/better score.

FBI: The pauses to allow scanning of questions and answers too short.

UM: I was able to answer because I circled the answers on the question sheets and then circled the appropriate letters on answer sheet.

UM: Instead of saying "Now answer quest. 1 it would be more productive to say "Who is Rodriguez" etc. (Some people are more audio than visual. You will be making more optimum use of tape.)

CAL: Too long at first, but then they right length later.

CAL: Too short -- add a few seconds. More time after the beep!
Test # ?
Have pp. 18-19 on facing pp.
When I took a moment to figure out a question, I would get caught trying to catch up with the recording. Pauses are fine for those who are proficient, but for semi-proficient examinees, they are too short.

CAL: 1) There seemed to be some background noise throughout the entire tape.
2) There are some discrepancies between tape/manual (noted in the form)

CAL: Too long for test 1, seemed about right for test.

GM: Could have been a little longer, would be better to answer as many questions as possible during the conversation.

GM: I knew what I knew right away, so the pauses seemed long, but then again I think I failed this class anyway, so there you have it.

GM: After hearing the conversation not much time is needed for answering. You either know the answer or not. Leaving too much time gives the answer too much time to think about it.

GM: However the pauses between like dialogs could have been slightly longer.

GM: Just give a few minutes to answer questions, instead of saying answer no. 1, pause, no. 2, pause, etc.

GM: A single block of time should be allowed to answer a group of questions, because some can be answered more quickly than others.

OTHER: They could have been a little shorter.

I thought they should cut out announcement of the numbers and say that there would be a minute or two to address all questions.

On Form 2 the directions say not to take notes and on Form 1 the directions say not to take notes or write in the booklet. Why the difference? I hadn't noticed that instruction and had crossed off or circled answers as I heard the conversation. This made it much easier. I think notes or marks on the Q's & A's test Spanish skills more and de-emphasize short-term memory skills.

PART B (SUMMARY)

General Comments:

FBI: This part was the difficult part for me!

1. **Were the directions for Part B clear?**

FBI: No problem.

FBI: The time in between seemed too long.

UM: Why did names have to be capitalized? Personal only? Countries? Organizations?

CAL: I guess it was hard to decide what information needed to be in. I think I gave a near and verbatim account.

GM: Very clear.

GM: No problems.

GM: Noteform is the most you could expect. Some names and places.

OTHER: Slow instructions!

2. **Was the example summary helpful?**

FBI: Good examples.

UM: There was some question about turning the page. We were asked to turn p. 14 twice.

CAL: Provision for answering of example or instructions regarding the fact that it is not necessary to answer them needs to be made clear.

CAL: Very much.

GM: There was no example summary!

3. **Do you have any comments about any of the conversations in Part B?**

FBI: Again, their dialect made it difficult.

FBI: The speaker's provided too much information for the time allotted to write down the translation.

FBI: 3rd conversation -- was too long. My hand is tired! And I got tired while writing.

FBI: Yes, the last conversation was somewhat difficult for me to understand and then translate. I didn't have enough time.

UM: Realistic and representative of actual situations which occur in our society.

UM: I have problems listening to taped conversations. Also, I think people should write their summary in Spanish. It is difficult translating in your head.

UM: I was glad that I was able to listen to conversation twice.

CAL: Good and clear. Form 2 was harder than Form 1 -- a more difficult topic and vocabulary and harder to understand the Spanish.

CAL: I think there was only male-male conversation. That is good, since those are very difficult to keep straight.

In both parts I had big problems remembering which voices belonged to which names. This was not problem when a male and female were talking, but caused big-time confusion when two men were talking. This is partly the result of low proficiency, but also the result of poor ability to distinguish two voices that are very similar. Do you want to test this ability.

CAL: Overall they seem well done.

Verbal comment: The third summary was really long on Part I (countries).

CAL: These are quite clear, which is better than many in Part A.

GM: They seemed hard, but some of it can understandable.

GM: Muy interesante.

GM: If you were listening to people in a real setup, you might have a frame of reference to start from and that would help.

GM: Again, kinda fast.

GM: Just that the last one (conversation 3) was too long. You lost the conversation about half way through.

GM: My comment would be the same as PART A -- with the little experience I have in interpreting conversations of a different language -- I found this section difficult as well.

GM: Some material was difficult to understand due to the fast speed of the dialog and being unfamiliar with some of the subject material.

GM: The last conversation was too long.

GM: Hard to summarize when can only catch key phrases or words. If we could just write words we understand and not have to tie it all together.

GM: Verv long and complicated.

GM: Very rapid.

OTHER: Easier to understand.

Los diálogos son más apropriados para todo nivel de estudiantes y sus futuras actividades en español.

Better than A but it was not made clear who spoke. Too political and drug oriented. More social conversation needed.

4. **Were the pauses for you to write the summaries about the right length?**

FBI: I think you should allow more time to write the summaries.

UM: About right -- except last one.

CAL: Too short for 3, about right for 1 and 2.

CAL. They might be too short for someone who understands more than I did. I couldn't write fast enough to include everything that I understood.

CAL: About right -- a bit too short for the last ones.

GM: No problems. Hard to catch names.

GM: Summaries that's too much. Besides in surveillance you would no doubt record. Perhaps tape should be stopped and started more like a translation exercise.

APPENDIX K

LSTE-SPANISH EXAM FORMAT

(FINAL VERSION)

FINAL VERSION

SPANISH LISTENING SUMMARY TRANSLATION EXAM FORMAT

Multiple Choice Items

TOPIC	FORM 1	TIME	FORM 2	TIME
Example	26[1] mf[2]	30s	26 mf	30s
General	27 mm	30s	28 mm	28s
Terrorism	31 mm	18s	32 mm	18s
Drugs	3 mf	49s	1a mf	1m 38s
	6a mf	1m 53s	1b mf	1m 28s
	6b mf	1m 14s	2a mm	1m
	8 mm	1m 50s	2c mf	51s
			4 mm	42s
Fraud	14b mm	2m 30s	9 mf	2m 33s
Immigration	16 mf	1m 45s	15 mm	1m 18s

Multiple Choice Section Subtotals

# of conversations (including one example)	8	9
Time	11m 19s	10m 48s
Interlocutors	4 mm 5 mf	5 mm 5 mf
# of multiple choice items developed	40	40

[1] Each scenario was assigned a number when originally drafted.

[2] Sex of interlocutors: mm = two males; mf = a male and a female

Summary Items

TOPIC	FORM A		TIME	FORM B		TIME
Example	25	mf	1m 10s	25	mf	1m 10s
General	30	mm	50s	29	mf	47s
Terrorism	17	mm	2m 6s	18	mm	2m 3s
FCI	23	mm	3m 10s	24	mm	3m 13s

Summary Section Subtotals

# of conversations (including one example)	4	4
Time	7m 16s	7m 13s
Interlocutors	3 mm 1 mf	2 mm 2 mf

Exam Totals

# of conversations	13	14
Time	18m 36s	18m 1s
Interlocutors	7 mm 6 mf	7 mm 7 mf

APPENDIX L

TEST ADMINISTRATION INSTRUCTIONS

TEST ADMINISTRATION INSTRUCTIONS

SPANISH LISTENING SUMMARY TRANSLATION EXAM

NOTE TO TEST ADMINISTRATOR

This manual describes important information about the procedures that must be followed BEFORE, DURING, and AFTER the administration of the translation exams. Uniform procedures are essential for the translation exams to yield reliable test results. The scores of all examinees from various field offices in the nation will be comparable only if all test administrators follow the same procedures and give exactly the same instructions. It is necessary, therefore, that you read the entire manual before administering the exams and follow the instructions without exception when administering the exams.

GENERAL INFORMATION

Test Security

It is extremely important that the translation exams be safeguarded and administered under secure conditions at each field office. In order to ensure test security, it is essential that you adhere to the following conditions:

1. Keep all test materials either in your immediate physical possession or in a locked cabinet or other secure area under your control.

2. Do not copy, or allow others to copy, any portion of the test booklets or tape, or make any notes or transcriptions of the test booklets or tape content.

3. Allow only those particular individuals who are to be tested to see the test materials, and only at the time of test administration and under the specific procedures described in this manual.

4. Should any irregularities occur, report them on the Test Administrator Report Form included in the test package. Please complete and sign this form even if no irregularities occur.

PRIOR TO THE TESTING DATE

Assembling Test Materials

Assemble as many test booklets and answer sheets as will be needed for the test administration, including an extra copy of each. You should also have on hand at least two no. 2 pencils (with erasers) for each examinee. Listed below are the materials needed for the Listening Summary Translation Exam:

1) Multiple Choice Section test booklets
2) Summary Section test booklets
3) Answer sheets
4) No. 2 pencils
5) Two copies of the tape for each form
6) A high-quality cassette playback unit

Arranging for a Testing Site

Locate a testing site that is comfortable and free from distraction. The listening exam requires a quiet room with good acoustics throughout and a high-quality cassette playback unit. The testing room should be large enough so that examinees can be seated with three feet of space in all directions between all examinees.

ON THE TESTING DATE

Equipment

Check the playback equipment to make sure that it is functioning properly. Adjust the volume control so that everybody in the room can hear the recording clearly. If the playback unit has a tone control, it should be set to the middle ("flat response") position or adjusted somewhat toward the treble. It should not be turned toward the bass position. Make sure that the tape is completely rewound after making these adjustments. Be sure to have two copies of the test tape on hand in case of a malfunction.

Prohibited Materials

Examinees may not use dictionaries during the Multiple Choice Section; however, they may use them during the Summary Section.

Administering the Test

Follow the procedures below when administering the test. All instructions within the boxes should be read verbatim. Do not depart from these directions unless noted otherwise.

1. After all have been seated, inform the examinees:

> **The Listening Summary Exam lasts approximately one hour and twenty minutes. All of the instructions for filling out the answer sheets are given on the test tape. There will be an opportunity to ask questions before the actual test begins.**

2. Distribute the Multiple Choice test booklets, answer sheets, and pencils.

3. Give the following instructions:

> **Please do not open your test booklet. In this section of the exam, you may mark your answers in the test booklet and then transfer them to the answer sheet. You must use a no. 2 pencil for marking your answers.**

4. Begin playing the tape.

5. Make sure that the form number given on the tape is the same as that of the

test booklets you have distributed.

6. Walk around the room to make sure that everyone is filling out the answer sheet correctly.

8. At the end of the Multiple Choice Section, inform the examinees:

> **This is the end of the Multiple Choice Section. Please stop working now. Now look over your answer sheet carefully. Be sure all the marks you made are dark and heavy. Insert your answer sheet in your test booklet and close your booklet.**

9. Immediately collect the Multiple Choice test booklets and answer sheets.

11. Distribute the Summary test booklets.

12. Fast forward the tape to the end and turn it over. Begin playing side two. (All of the instructions for the Summary Section are given on the tape.)

13. At the end of the test, inform examinees:

> **Please stop working now. Close your test booklets.**

14. Immediately collect the test booklets for the Summary Section. Be sure to account for all test booklets distributed. When all booklets have been accounted for, dismiss the examinees, or allow them to take a break before beginning the next exam.

APPENDIX M

SELF-ASSESSMENT OF SUMMARY ABILITY

NAME ____________________ FIELD OFFICE ________________

SELF-ASSESSMENT OF SUMMARY ABILITY

The purpose of this questionnaire is to learn your candid estimation of your ability to summarize in English conversations spoken in Spanish. It is of the utmost importance that you provide an honest evaluation of your present abilities so that the effectiveness of the summary exams may be accurately and fully assessed. Please be assured that your responses will be kept confidential by the test development contractor and will <u>in no way</u> affect your standing or possibility of advancement within the Bureau.

<u>Instructions</u>: FBI work involves the monitoring of conversations relating to narcotics trafficking, theft, white collar crime, organized crime, terrorism, and foreign counter-intelligence. These conversations vary in content, type of language (standard vs. highly colloquial), and style (direct vs. indirect). FBI employees are frequently called upon to summarize the information exchanged in the conversations. Please estimate your ability to summarize the following types of conversations using the scale provided below:

Limited — I can correctly report about half of the key points; however, I may not accurately represent the overall topic of the conversation.

Functional — I can correctly identify the topic of the conversation; however, my summary may contain misinterpretation or omission of several key points.

Competent — I can correctly report the topic of the conversation and most key and supporting points.

Superior — I can correctly report all key points and a wealth of supporting details including nuances of tone and emotion when appropriate.

Please evaluate candidly your ability to summarize the different types of conversations described below by circling the appropriate label:

<u>Type 1</u> — In Type 1 conversations, speakers generally use standard Spanish to communicate concrete information (dates, times, locations, amounts, etc.) in a direct manner.

Limited **Functional** **Competent** **Superior**

<u>Type 2</u> — In Type 2 conversations, speakers use a great deal of colloquial language (slang and regionalisms) to communicate concrete information (as above) in a fairly direct manner.

Limited **Functional** **Competent** **Superior**

<u>Type 3</u> — In Type 3 conversations, speakers use standard Spanish, possibly with colloquialisms, and make veiled or ambiguous references to shared knowledge (for example, "We'll meet tomorrow at the same place at the same time"); consequently, very little concrete information may be communicated.

Limited **Functional** **Competent** **Superior**

APPENDIX N

LSTE-SPANISH EXAM FEEDBACK QUESTIONNAIRE

(VALIDATION STUDY)

SPANISH LISTENING SUMMARY TRANSLATION EXAM

We would very much appreciate your answers to the following brief questions concerning the listening exam you have just taken:

1. Were the pauses for scanning the questions before the conversations about the right length?

   ( ) Too short
   ( ) About right
   ( ) Too long

2. Were the pauses for marking your answers on the answer sheet about the right length?

   ( ) Too short
   ( ) About right
   ( ) Too long

3. Were the pauses for writing the summaries about the right length?

   ( ) Too short
   ( ) About right
   ( ) Too long

Please indicate to what extent you agree or disagree with the following statements:

4. The directions for the multiple choice items were clear.

   ( ) Agree ( ) Disagree

5. The directions for the summary items were clear.

   ( ) Agree ( ) Disagree

6. The conversations (in both Parts A and B) were representative of the types of conversations I might encounter in my work.

   ( ) Strongly agree ( ) Agree ( ) Disagree ( ) Strongly disagree

7. There was sufficient opportunity for me to demonstrate my ability to understand and summarize conversations spoken in Spanish.

   ( ) Strongly agree ( ) Agree ( ) Disagree ( ) Strongly disagree

Thank you for your cooperation.

APPENDIX O

TEST ADMINISTRATOR REPORT FORM

Test Administrator Report Form

LISTENING SUMMARY EXAM

This form is to be used to report any irregularities in test administration. Please fill it out (even if there were no irregularities), sign your name, and return it with the test materials. Thank you.

* * * * * * * * *

Test Security

By agreeing to serve as the test administrator, I am responsible for ensuring the security of the test. I have kept the test materials confidential and secure at all times. None of the test booklets or test tapes has been reproduced in any form.

Irregularities: ____

____

Test Administration

The tests were administered in exact accordance with the procedures described in the Administration Manual. Any deviations from the stated procedures are listed below:

Irregularities: ____

____

Condition of Test Materials

Before returning the test materials, I have checked the condition of the test booklets and test tapes. All materials are being returned in their original condition.

Irregularities: ____

____

____

(Please print name) Field Office

____

Signature Date

APPENDIX P

LSTE-SPANISH EXAM FEEDBACK QUESTIONNAIRE RESULTS

(VALIDATION STUDY)

VALIDATION STUDY QUESTIONNAIRE

SPANISH LISTENING SUMMARY TRANSLATION EXAM

RESULTS

We would very much appreciate your answers to the following brief questions concerning the listening exam you have just taken:

1. Were the pauses for scanning the questions before the conversations about the right length?

(49%) Too short
(51%) About right
(0%) Too long

2. Were the pauses for marking your answers on the answer sheet about the right length?

(36%) Too short
(62%) About right
(2%) Too long

3. Were the pauses for writing the summaries about the right length?

(32%) Too short
(68%) About right
(0%) Too long

Please indicate to what extent you agree or disagree with the following statements:

4. The directions for the multiple choice items were clear.

(96%) Agree (4%) Disagree

5. The directions for the summary items were clear.

(98%) Agree (2%) Disagree

6. The conversations (in both Parts A and B) were representative of the types of conversations I might encounter in my work.

(23%) Strongly agree (49%) Agree (13%) Disagree (15%) Strongly disagree

7. There was sufficient opportunity for me to demonstrate my ability to understand and summarize conversations spoken in Spanish.

(21%) Strongly agree (38%) Agree (32%) Disagree (8%) Strongly disagree

Thank you for your cooperation.

APPENDIX Q

SUMMARY ACCURACY SCALE SCORE ESTIMATED FROM SPANISH OPI AND SELF ASSESSMENT

SUMMARY ACCURACY SCALE (SAS) SCORE ESTIMATED FROM OPI (SPANISH) AND SELF ASSESSMENT

ID	SPANSPK OPI	SELF ASSESSMENT	SUMMARY ACCURACY SCALE
45	3.8	10	4.10
46	2.0	8	3.00
47	3.0	8	3.35
48	3.8	11	4.34
49	5.0	11	4.76
60	0.8	3	1.38
61	1.0	3	1.45
62	0.8	3	1.38
65	1.0	4	1.69
66	0.8	3	1.38
67	1.0	3	1.45
68	1.0	4	1.69
69	1.0	4	1.69
70	1.0	3	1.45
71	1.0	3	1.45
72	1.0	4	1.69
73	1.0	3	1.45
74	1.0	3	1.45
75	1.0	3	1.45
76	0.8	3	1.38
77	0.8	5	1.85
78	1.0	6	2.17

SUMMARY ACCURACY SCALE (SAS) SCORE ESTIMATED FROM
OPI (SPANISH) AND SELF ASSESSMENT

ID	SPANSPK OPI	SELF ASSESSMENT	SUMMARY ACCURACY SCALE
1	5.0	10	4.52
2	4.8	10	4.45
3	3.8	6	3.15
4	3.0	10	3.82
5	4.8	10	4.45
6	2.0	7	2.76
7	4.8	9	4.21
8	2.0	10	3.47
9	4.8	12	4.93
10	5.0	10	4.52
11	3.8	8	3.63
12	4.8	11	4.69
13	3.0	12	4.30
14	5.0	12	5.00
15	4.8	8	3.98
16	5.0	9	4.28
17	5.0	7	3.81
18	5.0	12	5.00
19	4.0	11	4.41
20	2.0	6	2.52
22	3.8	12	4.58
23	4.0	12	4.65
24	4.8	8	3.98
25	2.8	6	2.80
26	5.0	10	4.52
27	4.0	12	4.65
28	5.0	11	4.76
29	4.0	5	2.98
30	5.0	10	4.52
32	2.8	9	3.51
34	4.8	10	4.52
36	4.8	12	4.92
37	5.0	8	4.05
40	3.8	9	3.86

APPENDIX R

INTERPRETATION OF FINAL ACCURACY RATING

INTERPRETATION OF FINAL ACCURACY RATING

NO ABILITY
No response or fails to identify overall topic accurately. Typically provides no substantial information beyond names of speakers.

SEVERELY DEFICIENT
Often fails to identify topic accurately. Contains frequent misinterpretations, omissions, and/or misleading additions. Usually less than a fourth of the key points of information are correctly reported.

DEFICIENT
May not represent topic accurately. Contains many misinterpretations, omissions, and/or misleading additions. About half of the key points of information may be correctly reported.

FUNCTIONAL
Normally identifies topic accurately; however, contains misinterpretation, omission, and/or misleading addition of several key points of information. May contain a number of supporting details.

COMPETENT
Accurately reports almost all key points of information and many supporting details; no misleading additions.

SUPERIOR
Accurately reports all or almost all key points of information and supporting details.

APPENDIX S

RAW SCORE TO SUMMARY ACCURACY SCALE SCORE CONVERSION TABLES

Score Conversion Table - Form 1
Multiple Choice Section Score to Summary Accuracy Scale

Form 1 Multiple Choice Raw Score	Summary Accuracy Scale Score
1	*
2	*
3	*
4	*
5	*
6	*
7	*
8	*
9	*
10	*
11	.68
12	.83
13	.98
14	1.12
15	1.27
16	1.42
17	1.57
18	1.71
19	1.86
20	2.01
21	2.15
22	2.30
23	2.45
24	2.59
25	2.74
26	2.89
27	3.04
28	3.18
29	3.33
30	3.48
31	3.62
32	3.77
33	3.92
34	4.07
35	4.21
36	4.36
37	4.51
38	4.65
39	4.80
40	4.95

* 1-10 = chance scores

Regression Standard Error of Estimate = .83

Score Conversion Table - Form 2
Multiple Choice Section Score to Summary Accuracy Scale

Form 2 Multiple Choice Raw Score	Summary Accuracy Scale Score
1	*
2	*
3	*
4	*
5	*
6	*
7	*
8	*
9	*
10	*
11	1.21
12	1.34
13	1.47
14	1.60
15	1.72
16	1.85
17	1.98
18	2.11
19	2.24
20	2.37
21	2.50
22	2.62
23	2.75
24	2.88
25	3.01
26	3.14
27	3.27
28	3.39
29	3.52
30	3.65
31	3.78
32	3.91
33	4.04
34	4.16
35	4.29
36	4.42
37	4.55
38	4.68
39	4.81
40	4.93

* 1-10 = chance scores

Regression Standard Error of Estimate = .86

Score Conversion Table - Form 1
Total Accuracy Score to Summary Accuracy Scale

Total Accuracy	Sum Accuracy Scale	Total Accuracy	Sum Accuracy Scale	Total Accuracy	Sum Accuracy Scale
0	.04	37	1.72	74	3.39
1	.09	38	1.76	75	3.44
2	.13	39	1.81	76	3.48
3	.18	40	1.85	77	3.53
4	.22	41	1.90	78	3.57
5	.27	42	1.94	79	3.62
6	.31	43	1.99	80	3.66
7	.36	44	2.03	81	3.71
8	.40	45	2.08	82	3.75
9	.45	46	2.12	83	3.80
10	.50	47	2.17	84	3.85
11	.54	48	2.22	85	3.89
12	.59	49	2.26	86	3.94
13	.63	50	2.31	87	3.98
14	.68	51	2.35	88	4.03
15	.72	52	2.40	89	4.07
16	.77	53	2.44	90	4.12
17	.81	54	2.49	91	4.16
18	.86	55	2.53	92	4.21
19	.90	56	2.58	93	4.25
20	.95	57	2.62	94	4.30
21	.99	58	2.67	95	4.34
22	1.04	59	2.71	96	4.39
23	1.08	60	2.76	97	4.43
24	1.13	61	2.80	98	4.48
25	1.17	62	2.85	99	4.52
26	1.22	63	2.89	100	4.57
27	1.26	64	2.94	101	4.61
28	1.31	65	2.98	102	4.66
29	1.36	66	3.03	103	4.70
30	1.40	67	3.08	104	4.75
31	1.45	68	3.12	105	4.80
32	1.49	69	3.17	106	4.84
33	1.54	70	3.21	107	4.89
34	1.58	71	3.26	108	4.93
35	1.63	72	3.30	109	4.98
36	1.67	73	3.35	110-114	5.00

Regression Standard Error of Estimate = .67

Score Conversion Table - Form 2
Total Accuracy Score to Summary Accuracy Scale

Total Accuracy	Sum Accuracy Scale	Total Accuracy	Sum Accuracy Scale	Total Accuracy	Sum Accuracy Scale
0	.31	34	1.91	68	3.51
1	.36	35	1.96	69	3.55
2	.41	36	2.00	70	3.60
3	.45	37	2.05	71	3.65
4	.50	38	2.10	72	3.70
5	.55	39	2.15	73	3.74
6	.59	40	2.19	74	3.79
7	.64	41	2.24	75	3.84
8	.69	42	2.29	76	3.88
9	.74	43	2.33	77	3.93
10	.78	44	2.38	78	3.98
11	.83	45	2.43	79	4.02
12	.88	46	2.47	80	4.07
13	.92	47	2.52	81	4.12
14	.97	48	2.57	82	4.17
15	1.02	49	2.61	83	4.21
16	1.06	50	2.66	84	4.26
17	1.11	51	2.71	85	4.31
18	1.16	52	2.76	86	4.35
19	1.21	53	2.80	87	4.40
20	1.25	54	2.85	88	4.45
21	1.30	55	2.90	89	4.49
22	1.35	56	2.94	90	4.54
23	1.39	57	2.99	91	4.59
24	1.44	58	3.04	92	4.64
25	1.49	59	3.08	93	4.68
26	1.53	60	3.13	94	4.73
27	1.58	61	3.18	95	4.78
28	1.63	62	3.23	96	4.82
29	1.68	63	3.27	97	4.87
30	1.72	64	3.32	98	4.92
31	1.77	65	3.37	99	4.96
32	1.82	66	3.41	100-111	5.00
33	1.86	67	3.46		

Regression Standard Error of Estimate = .60